Single-molecule Spectroscopy of the Folding Dynamics and Unfolded State Properties of a Fast Folding Protein

---

Dissertation
zur
Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)
vorgelegt der
Mathematisch-naturwissenschaftlichen Fakultät
der
Universität Zürich

von
René Wuttke genannt Krause
aus
Deutschland

Promotionskomitee
Prof. Dr. Benjamin Schuler (Vorsitz)
Prof. Dr. Raimund Dutzler
Prof. Dr. Peter Hamm

August 2012

Abstract

Understanding protein folding, the process in which polypeptide chains acquire their three-dimensional topology, is one of the major challenges in biology. Protein folds can differ greatly, although they all originate from a linear educt of relatively similar chemistry. To understand the folding reaction, it is important to investigate its starting point, the unfolded state of the protein, as well as its free-energy landscape. The use of single-molecule techniques holds the promise of accomplishing just that. It allows the study of the unfolded state in the presence of the native state and avoids ensemble averaging of the folding reaction, in principle making the whole energy landscape accessible to experimental observation. In this thesis I use single molecule Förster resonance energy transfer spectroscopy, which reports on intramolecular distances, to study several aspects of the unfolded state properties and the folding reaction.

In the first part of the thesis we investigate the temperature-induced collapse of unfolded proteins. Counterintuitively, unfolded proteins do not expand with increasing temperature like unstructured homopolymers in the physiologically relevant temperature range, but show a collapse. To elucidate the driving force behind this behavior, we compare different proteins with vastly different sequence compositions, varying in degree of charge content and hydrophobicity. We find that temperature-induced collapse seems to be a general property of proteins, and is not limited to particularly hydrophobic sequences. However, λ repressor, the most hydrophobic protein investigated, shows a strong collapse, followed by a slight re-expansion at higher temperatures. The reasons for this behavior are unclear, but they do not seem to involve a strong contribution of secondary structure formation. The relationship of the degree of collapse, which we represent as an intra-chain energy determined based on Sanchez theory, and the chain composition does not seem to be a simple linear correlation. This might point to more complex protein-solvent interactions than previously thought.

In the second part of the thesis we take first steps to resolve barrier crossing events in the folding and unfolding of single λ repressor molecules. We aim to do this in the unhindered environment of freely diffusing molecules, removing potential artifacts from surface tethering and allowing the observation of a large number of events. To be able to observe many (un)folding events, the protein needs to have a folding time at the unfolding midpoint in the range of the diffusion time through the confocal volume. We tried to engineer such a λ repressor variant and present data analysis methods to obtain information on the folding kinetics from the same single-molecule equilibrium experiments, most notably a maximum likelihood method developed by Gopich and Szabo and the recurrence analysis of single particles. Our analysis results provide valuable starting points for the study of ultrafast protein folding in free solution.

In the last part of this thesis, a theoretical approach developed by Philipp Schütz *et al.* is applied to experimental data of λ repressor. He proposed a new method to obtain the free energy landscape of the protein folding reaction from a time series of single distance information such as the one provided from the FRET efficiency. In the "Free Energy Surface from Single-Molecule Time Series" (FESST) method, an equilibrium transition network is constructed by clustering individual time windows of the distance information, in which then free energy basins are identified by a minimum-cut-based Free Energy Profile (cFEP) approach. The method underestimates the folding barrier, but successfully identifies folded and unfolded populations, and provides a promising new approach for the analysis of single molecule experiments.

Zusammenfassung

Proteinfaltung ist der Prozess, durch den Polypeptide ihre räumliche Struktur einnehmen, und stellt eines der wichtigsten ungelösten Probleme in der Biologie dar. Obwohl das Edukt der Proteinfaltungsreaktion - der ungefaltete Zustand - linear und chemisch relativ homogen ist, kann eine sehr grosse Vielfalt an Topologien erreicht werden. Um die Faltung zu verstehen, muss man sowohl diesen Ausgangszustand als auch die komplette Freie Energie-Landschaft der Reaktion verstehen. Einzelmolekülverfahren haben das Potential, genau das zu erreichen. Sie ermöglichen die Untersuchung des ungefalteten Zustands auch in Anwesenheit von nativem Protein und vermeiden Herausmittelung des eigentlich relevanten Signals individueller Faltungsreaktionen. Damit wäre theoretisch eine komplette Beschreibung der Energielandschaft möglich. In dieser Arbeit nutze ich Einzelmolekül-Förster-Resonanzenergietransferspektroskopie, mit der man intramolekulare Distanzen messen kann, um verschiedene Aspekte der Eigenschaften des ungefalteten Zustands und der Faltungsreaktion zu untersuchen.

Im ersten Teil der Arbeit untersuchen wir den temperaturinduzierten Kollaps des ungefalteten Zustands von Proteinen. Obwohl durch die Erhöhung der Temperatur die Entropie des System steigt, expandieren ungefaltete Proteine nicht wie unstrukturierte Homopolymere über die physiologisch relevante Temperaturspanne, sondern zeigen einen Kollaps. Um die treibende Kraft hinter diesem Effekt zu untersuchen, vergleichen wir verschiedene Proteine mit deutlich unterschiedlichen Aminosäuresammensetzungen in Bezug auf Ladung und Hydrophobizität. Dabei stellen wir fest, dass temperaturinduzierter Kollaps eine allgemeine Eigenschaft von Proteinen ist und nicht nur auf hydrophobe Sequenzen limitiert ist. Der λ Repressor, das am meisten hydrophobe untersuchte Protein, zeigt sowohl einen starken Kollaps als auch eine leichte Reexpansion bei höheren Temperaturen. Der Grund hierfür ist unklar. Es scheint sich jedoch nicht um einen starken Einfluss von Sekundärstruktur zu handeln. Die Beziehung zwischen dem Ausmass an Kollaps, welchen wir als interne Kettenenergie repräsentieren, die wir über die Sanchez-Theorie erhalten, und der Sequenzzusammensetzung ist kein einfacher linearer Zusammenhang. Dies deutet auf komplexere Protein-Lösungsmittel-Interaktionen hin als zuvor angenommen.

Im zweiten Teil der Arbeit machen wir erste Schritte in Richtung der zeitlichen Auflösung der Barrierenüberquerung in der Faltungsreaktion von einzelnen λ-Repressor-Molekülen. Um Artefakte durch eine Oberflächenimmobilisierung zu vermeiden und eine einfache Akkumulierung vieler Messdaten zu erleichtern, verwenden wir konfokale Detektion in Kombination mit Proteinen, die frei in Lösung diffundieren. Damit möglichst viele individuelle Reaktionen beobachtet werden können, ist es notwenig, dass das untersuchte Protein eine Faltungszeit am Entfaltungsmittelpunkt hat, die im Bereich der Diffusionszeit durch das konfokale Volumen liegt. Wir versuchten, eine geeignete λ-Repressor-Variante herzustellen

und präsentieren verschiedene Analyseverfahren zur Ermittlung von kinetischen Daten direkt aus Einzelmolekülmessungen im chemischen Gleichgewicht - insbesondere eine *Maximum Likelihood*-Methode von Gopich und Szabo sowie die *Recurrence Analysis of Single Particles* (RASP). Unsere Untersuchungen zeichnen bisher jedoch ein nicht eindeutiges Bild der λ Repressor-Faltung, bieten jedoch wichtige Anknüpfungspunkte für weitere Studien an ultraschnell faltenen Proteinen in Lösung.

Im letzten Teil der Arbeit wird ein theoretischer Ansatz, welcher von Philipp Schütz *et al.* entwickelt wurde, auf experimentelle Daten von λ-Repressor angewendet. Er schlug eine neue Methode vor, um die Freie Energie-Landschaft der Proteinfaltung lediglich aus einer einzelnen Distanzinformation zu erhalten. Bei der FESST-(*Free Energy Surface from Single-molecule Time Series*) Methode wird ein Gleichgewichtsnetzwerk von Übergängen aus kurzen Abschnitten der Distanzinformation konstruiert. In diesem Netzwerk werden Minima der Freien Energie mit dem *Minimum-Cut-Based Free Energy Profile* (cFEP) -Ansatz identifiziert. Die Methode unterschätzt dabei die Faltungsbarriere, aber identifiziert erfolgreich die gefalteten und ungefalteten Populationen, und stellt somit einen vielversprechenden neuen Ansatz in der Auswertung von Einzelmolekülexperimenten dar.

Contents

Chapter 1

Introduction

1.1 Protein folding

The existence of proteins is one of the reasons that life as we know it exists. Proteins are Nature's robots that control, partition and transform all other components and reactions in the metastable system of the living organism. They achieve this breadth of function via huge structural diversity. For example, approximately 35,000 kinds of proteins are found in the human body alone. Though post-translational modification and alternative splicing of underlying mRNA sequences play an important role, the main source of diversity of the proteome is the primary sequence of a protein: the order in which the building blocks – the 20 proteinogenic amino acids - are arranged in a linear fashion. Anfinsen postulated (Anfinsen, 1973) that this information, together with the solution conditions, is enough for a peptide to assume a unique and thermodynamically stable conformation, in contrast to the idea of a mechanism in which the population of a native state is dependent on the folding route. While it has been shown that many proteins need additional factors such as chaperones to fold *in vivo* (Thirumalai and Lorimer, 2001), Anfinsen's dogma holds for almost all proteins investigated so far. In retrospect, this might not be surprising, as proteins have evolved to perform a single or multiple distinct functions, a common prerequisite of which is that they attain a certain conformation to be active (although with the discovery of intrinsically disordered proteins this is not true for every protein (Dyson and Wright, 2005)). One would suspect it to be beneficial that this conformation would be attained in an accurate and timely fashion.

However, from a purely statistical point of view, it is a miracle that any protein ever gets folded at all. Levinthal proposed the following thought experiment (Levinthal, 1969): Consider a 100 amino acid polypeptide in which every residue could assume only one of two possible states (2^{100} degrees of freedom) and the interconversion timescale of chain conformations is in the subpicosecond regime (10^{-13} sec). If the protein was to sample every possible

conformation until reaching the native state, it would take, $4 \cdot 10^9$ years, the approximate age of Earth (Dalrymple, 2001). In reality, most proteins do not take more than a few minutes to fold and many reach their final native state in mere microseconds. These effects are easier to understand if one considers one of the most popular theories to understand protein folding – energy landscape theory, proposed by Bryngelson and Wolynes in the 1980s and 1990s (Bryngelson and Wolynes, 1987; Bryngelson *et al.*, 1995; Dill and Chan, 1997; Onuchic *et al.*, 1997; Pande *et al.*, 1998; Dinner *et al.*, 2000; Plotkin and Onuchic, 2002). It depicts how the free energy of a reacting system depends on its degrees of freedom. For protein folding reactions, these degrees of freedom, can be anything from the distance between two specific atoms or the number of native contacts, which might be only accessible in all-atom MD simulations (Scheraga *et al.*, 2007), to more practical measures as chain end-to-end distances (Rischel and Poulsen, 1995; Schuler *et al.*, 2002; Möglich *et al.*, 2006; Schuler, 2007; Sherman *et al.*, 2008) or hydrodynamic radii (Wilkins *et al.*, 1999; Wu *et al.*, 2008). In this description, the native state can be found at a set of coordinates on this landscape (see figure 1.1). The remaining ensemble of conformations comprises the non-native ensemble, which is made up of unstructured conformations, intermediates, misfolded states and transition states. Depending on the free energy of a certain conformation, it will be differently populated. Of particular interest is the well of the unfolded state, as discussed in the following section. The diffusive process of moving along the lowest free energy route from the unfolded ensemble well to the native state is what is considered the folding of the protein. The energy profile along this pathway will determine the kinetics of the folding reaction. Whilst the field of protein folding has its roots in the mid 20$^{\text{th}}$ century, the questions raised remain relevant and experimentally and theoretically challenging. The behavior of a protein on the multi-dimensional landscape provides a wealth of information, some aspects of which this thesis aims to elucidate *via* a combination of experiment and application of theory.

1.2 The unfolded state of proteins

Whilst the native structure of a protein is essential for its function, a moderately stable protein with typical folding kinetics will cycle through several unfolding and folding events during its *in vivo* lifetime. Although particularly stable proteins do exist (Jaenicke, 1996), most proteins have to visit the denatured state on several occasions. These include translation at the ribosome, translocation across membranes, denaturing stress conditions of the organism and finally prior to the proteins degradation. Furthermore, the class of natively unfolded proteins functions in partially or completely unfolded states. In addition to the unfolded state's functional role and its potential to partake in *in vivo* interactions, the unfolded state is also

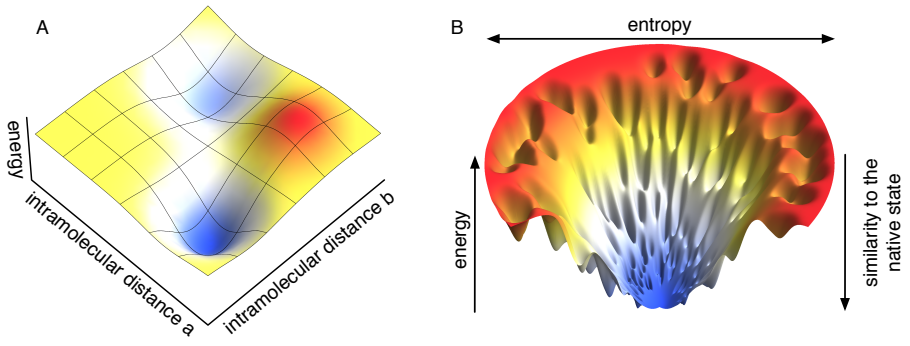

Figure 1.1: Two common representations of free energy landscapes. **A** A schematic representation of a two-dimensional energy landscape, plotting the free energy of the chain against two intramolecular spatial measures. The latter can be e.g. particular dihedral angles or atom-to-atom distances but also other configurational values. The unfolded and the native states are represented as wells. Folding pathways are all connections between the two. The saddle point on a minimal energy path between the two wells corresponds to the transition state. **B** A second common depiction is the funnel-like energy landscape. The depth of the funnel here represents the effective energy. The width corresponds to the entropy of the configurational ensemble. Folding is a diffusive process down the funnel to the unique native structure. Misfolded conformations correspond to bumps in the wall of the funnel.

interesting from the standpoint of kinetics as the starting point of the folding reaction and as a factor determining protein stability. To understand these folding kinetics, one has to understand the unfolded state's properties under different conditions first. How do these change due to different solvent conditions, for example concentration of denaturant, temperatures, ion strength, viscosity or pressure?

For a long time, it has been attempted to describe the behavior of the unfolded state in terms of a theoretical framework (Flory, 1945; De Gennes, 1975; Sanchez, 1979; Chan and Dill, 1991; Dill and Shortle, 1991). One can consider proteins as chains of monomers. These monomers can interact with each other (they obviously have to, to form the native state), but cannot occupy the same volume in space (as they would in a random walk model) – this is known as the effect of excluded volume. Also, they can interact with the surrounding solvent. The relative partitioning between intrachain and chain-solvent interactions is determined by the solution conditions listed above. This will decide if solvents are considered "good", "theta" or "poor" for a certain chain. In good solvents, monomers interact favorably in terms of free energy with the solvent, leading to expansion, on the other hand in poor solvents intrachain interactions are preferred, resulting in chain collapse as the solvent-chain interfacial area is minimized. Theta solvents lie in between both cases, where the solvation is just so unfavorable that it cancels out the chain expansion by excluded volume effects. How to describe the inho-

mogeneous ensemble of the denatured state conformations within a theoretical framework has been, and remains, a heavily debated issue. Additionally, one has to take into consideration that proteins have a heterogeneous sequence of monomers. Rather than the more simplified model discussed previously, each amino acid imposes its own charge and spatial constraints on the system, the strength of which is tuned by different solvent conditions by varying extents. This leads to the open question, whether specific interactions are present in the unfolded state and therefore how structured it is. Evidence for native-like and non-native interactions has been found using different experimental approaches. One approach is to directly measure characteristics of residual or secondary structure, e.g. by probing local interactions by NMR (Shortle and Ackerman, 2001; Klein-Seetharaman *et al.*, 2002) or measuring secondary structure content by circular dichorism (Yang *et al.*, 2003; Hoffmann *et al.*, 2007). Another approach is to measure overall chain dimensions by overall hydrodynamic radii or long-range distance information, and compare them with the dimensions expected from the length-scaling of different polymer models, like the random flight or the self-avoiding chain. This has been the strategy of scattering methods (Millett *et al.*, 2002; Gast and Modler, 2008). However, reasonable doubt has arisen that derivations from the scaling behavior of a Gaussian chain are necessarily indicative of residual structure, if enough flexible sections of the chain connect ordered parts (Fitzkee and Rose, 2004). The fact that the unfolded state is a rapidly interconverting ensemble of conformations severely complicates the experimental investigation of its properties. Residual structure might not be present in all parts of the protein at all times, emphasizing the need for a high spatiotemporal resolution. Furthermore, especially ensemble methods struggle to investigate the unfolded state under native conditions, where it is of the highest importance. The interpolation of biophysical characteristics of the unfolded state from experiments in high concentrations of denaturant back to native conditions might not be trivial or the alternative approach of sufficiently populating the unfolded state through destabilization of the protein by mutation potentially skews the result. So in addition to the averaging of the signal over the unfolded conformations, one has to account for a contribution of the folded state in such experiments.

1.3 Single-molecule FRET

Single molecule techniques inherently have several advantages over ensemble approaches. Most importantly, single molecule techniques avoid signal averaging over multiple conformational states, allowing the study of heterogeneous mixtures. In this study, it sheds light on the protein unfolded state under folding competent conditions. Since proteins can be found at many positions on the energy landscape, single molecule experiments can, in theory, explore

it in full and in doing so inform on the specific pathway of every folding reaction individually. However, this is limited by the difficulty of obtaining enough data from low populated states in order to satisfy statistical rigor and the sampling rate of the particular single molecule experiment in relation to the timescale of the folding process. Furthermore single molecule techniques have the added benefit that, due to working at very low concentrations, protein aggregation is generally avoided. This enables the study of proteins in solution conditions where higher protein concentrations would lead to aggregation, which in turn would at the very least influence the observed signal or make certain regions of the folding landscape kinetically inaccessible. Although this has not been an issue with λ repressor.

While information about single molecules can also be obtained by force probe techniques like atomic force microscopy (Rief *et al.*, 1997; Borgia *et al.*, 2008) and optical tweezers (Neuman and Block, 2004), we employed Förster resonance energy transfer (FRET) (Förster, 1948; Stryer and Haugland, 1967; Ha *et al.*, 1996; Jia *et al.*, 1999; Deniz *et al.*, 2000; Moerner, 2002; Schuler and Eaton, 2008). In this approach, two distinct fluorophores are chemically attached to the protein. A donor fluorophore will absorb excitation light and then will, from its excited state, either emit fluorescence light or transfer its energy in a radiation-free dipole-dipole coupling to the acceptor fluorophore. The efficiency of the transfer process depends on several factors inherent to the choice of experimental conditions and choice of donor-acceptor pair, and most importantly on the distance between the fluorophores. This enables FRET to serve as a "spectroscopic ruler" (Stryer and Haugland, 1967; Schuler *et al.*, 2005). These relations can be formalized according to Försters equation:

$$E = \frac{R_0^6}{R_0^6 + r^6} \tag{1.1}$$

The transfer efficiency E depends on interfluorophore distance, r, by the inverse sixth power as illustrated in figure 1.2. The Förster radius (Förster, 1948), R_0, which is dependent on the type of fluorophores, denotes the distance at which the transfer efficiency is 50%, which in turn depends on several factors and can be calculated as:

$$R_0^6 = \frac{9000(\ln 10)\kappa^2 \phi_D J}{128\pi^5 n^4 N_A} \tag{1.2}$$

where κ^2 is the orientation factor of the dye transition dipoles, ϕ_D is the donor quantum yield, J is the overlap integral between donor emission and acceptor absorption spectra , n is the refractive index of the medium between the dipoles, and N_A is Avogardo's number. In the case of chromophores that can rotate freely and significantly faster than the timescale of the fluorescence life time, $\kappa^2 = 2/3$ can be assumed. This approximation holds for most dye pairs attached to proteins in the native state and is nearly always true in the presence of denaturants such as GdmCl (dos Remedios and Moens, 1995; Allen and Paci, 2009). The solution condi-

tions (buffer and temperature) influence n , which can be measured directly. With common dye pairs, this leads to values of R_0 that enable an optimal window of sensitivity between 2 and 9 nm, which is compatible with the typical lengthscales in biomolecules and protein-protein interactions. To gain access to the intramolecular distance distribution underlying the

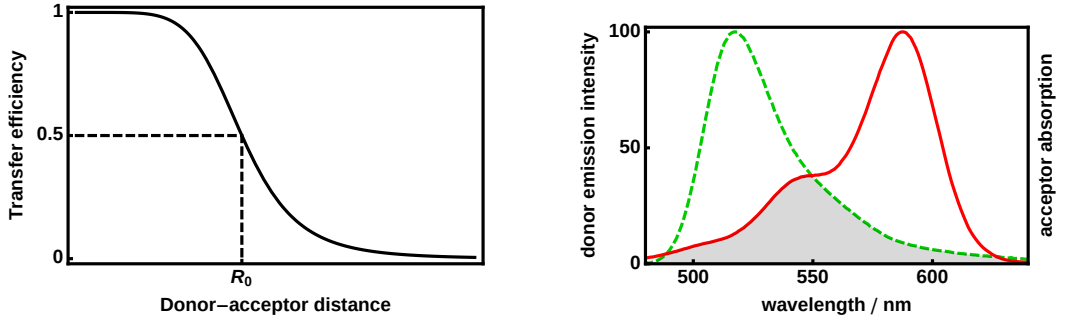

Figure 1.2: Förster Energy Transfer Left: Distance dependence of the transfer efficiency according to Förster's equation (equation 1.1), illustrating also that the sensitivity of FRET experiments to distance changes between the chromophores is greatest around the characteristic Förster distance, but quickly falls off at shorter or longer distances. This is amongst other factors largely dependent on the choice of fluorophores. Right: Donor's emission and acceptors absorption spectrum for the employed dye pair of Alexa 488 (donor, green) and 594 (acceptor, red). The gray area corresponds to the overlap of both spectra, which greatly influences the the value of R_0.

transfer efficiency distribution, one can measure the latter in different ways. The most common is based on fluorescence intensities, or rather individual photons in the context of single molecule detection. Donor fluorophores will emit fluorescence light at shorter wavelengths than the acceptor, enabling their spectral separation.

$$E = \frac{n'_A}{n'_A + n'_D} \tag{1.3}$$

n'_D and n'_A are the number of photons in the donor and acceptor channels, respectively. These, however, need to be corrected for the optical and detection setup of the instrumentation as well as the used dye pair (as outlined in Materials and Methods). A complementary approach is the utilization of fluorescence lifetimes:

$$E = 1 - \frac{\tau_{DA}}{\tau_D} \tag{1.4}$$

where τ_{DA} is the donor lifetime in the presence of the acceptor, while τ_D is the donor lifetime in the absence of the acceptor. The instrumentation is outlined in figure 1.3. It is based on a microscope with confocal excitation and detection and single-photon detection capabilities. The single molecule detection is achieved by focussing into a very small detection volume in the femtolitre regime, and sample concentrations in the range of 10-100 pM. Therefore the probability of having more than one molecule in the focus is extremely low.

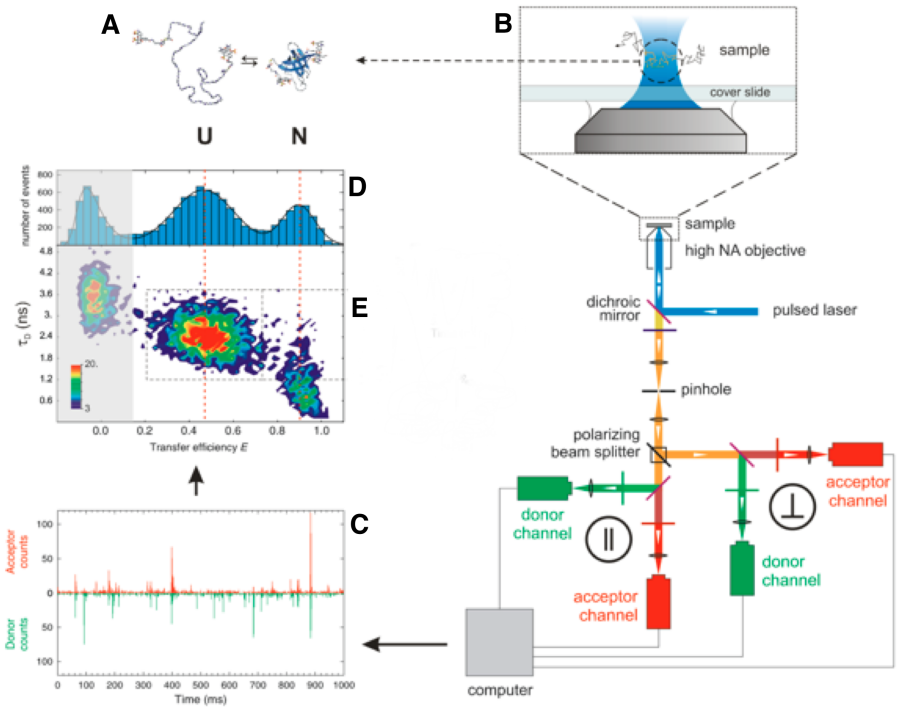

Figure 1.3: Typical scheme of confocal single-molecule FRET experiments Molecules at picomolar concentrations diffuse in different conformations (A) through the confocal volume (B). Laser excitation and collection of fluorescence light is performed through the same objective. The later is split according to its polarization and color. Single photons are (time tagged) detected and analyzed in trajectories (C). Individual molecules are identified by the photon bursts they cause, and are analyzed in terms of transfer efficiencies (D) or additionally other parameters, e.g. fluorescence lifetime (E).(adapted from Schuler 2005)

1.4 Aims

In this thesis I aimed to shed some light on current problems in the study of proteins, their folding and how to apply single-molecule FRET methods to answer them. The microscopic pathways of the folding reaction on the energy landscape are complex, but in the context of single-molecule FRET one usually has only one observable available, the transfer efficiency E. In addition to this, the observation time, signal-to-noise-ratio , or the difference in E of different conformational state are often limited. This calls for an adequate analysis method adapted to these limitations. Baba and Komatsuzaki (2007) proposed such a method, an extended version of which we apply here for the first time to single molecule FRET data. This also holds the promise of a more direct comparison of experimental and molecular dynamic simulation data.

How the protein unfolded states respond to changing solvent conditions, such as the additions of denaturants (Hoffmann *et al.*, 2007; Ziv and Haran, 2009) has been the topic of many studies. In recent years, the influence of the temperature has come into focus (Yang *et al.*, 2003; Nettels *et al.*, 2009). A collapse of the unfolded state upon temperature increase has been shown. This is contrary to the expansion expected for an unstructured homopolymer. The hydrophobic effect is thought to increase in strength with increasing temperature, also it is believed that secondary structure might emerge, both potentially resulting in a collapse. Since the primary sequence of a protein will affect both properties, we present here a comparative study on how the unfolded states of different proteins respond to temperature.

Finally, we set out to observe single molecules during their folding reaction, with the goal to resolve the process of crossing the free energy barrier. This ultimately holds the promise of a complete reconstruction of the energy landscape of the protein. However, this experiment is complicated by the fast timescale of the folding reaction and the low time resolution and short observation time of current single-molecule FRET experiments.

Bibliography

Allen L.R.; Paci E. Orientational averaging of dye molecules attached to proteins in Förster resonance energy transfer measurements: insights from a simulation study. *The Journal of Chemical Physics*, **131**(6):065101 (2009).

Anfinsen C.B. Principles that govern the folding of protein chains. *Science*, **181**(4096):223–230 (1973).

Baba A.; Komatsuzaki T. Construction of effective free energy landscape from single-molecule

time series. *Proceedings of the National Academy of Sciences of the United States of America*, **104**(49):19297–19302 (2007).

Borgia A.; Williams P.M.; Clarke J. Single-molecule studies of protein folding. *Annual review of biochemistry*, **77**:101–125 (2008).

Bryngelson J.D.; Onuchic J.N.; Socci N.; Wolynes P. Funnels, Pathways and the Energy Landscape of Protein Folding: A Synthesis. *Proteins: Structure, Function, and Bioinformatics*, **21**(3):167–195 (1995).

Bryngelson J.D.; Wolynes P.G. Spin glasses and the statistical mechanics of protein folding. *Proceedings of the National Academy of Sciences of the United States of America*, **84**(21):7524–7528 (1987).

Chan H.S.; Dill K.A. Polymer principles in protein structure and stability. *Annual review of biophysics and biophysical chemistry*, **20**:447–490 (1991).

Dalrymple G.B. The age of the Earth in the twentieth century: a problem (mostly) solved. *Special Publications, Geological Society of London*, **190**(1):205–221 (2001).

De Gennes P. Collapse of a polymer chain in poor solvents. *Journal de Physique Lettres* (1975).

Deniz A.; Laurence T.; Beligere G.; Dahan M.; Martin, AB; Chemla D.; Dawson P.; Schultz P.; Weiss S. Single-molecule protein folding: Diffusion fluorescence resonance energy transfer studies of the denaturation of chymotrypsin inhibitor 2. *Proceedings of the National Academy of Sciences of the United States of America*, **97**(10):5179–5184 (2000).

Dill K.; Chan H. From Levinthal to pathways to funnels. *Nature Structural Biology*, **4**(1):10–19 (1997).

Dill K.A.; Shortle D. Denatured states of proteins. *Annual review of biochemistry*, **60**:795–825 (1991).

Dinner A.R.; Sali A.; Smith L.J.; Dobson C.M.; Karplus M. Understanding protein folding via free-energy surfaces from theory and experiment. *Trends in biochemical sciences*, **25**(7):331–339 (2000).

dos Remedios C.G.; Moens P.D. Fluorescence resonance energy transfer spectroscopy is a reliable "ruler" for measuring structural changes in proteins. Dispelling the problem of the unknown orientation factor. *Journal of structural biology*, **115**(2):175–185 (1995).

Dyson H.J.; Wright P.E. Intrinsically unstructured proteins and their functions. *Nature reviews Molecular cell biology*, **6**(3):197–208 (2005).

Fitzkee N.C.; Rose G.D. Reassessing random-coil statistics in unfolded proteins. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(34):12497–12502 (2004).

Flory P. Thermodynamics of Dilute Solutions of High Polymers. *The Journal of Chemical Physics*, **13**(11):453–465 (1945).

Förster T. Zwischenmolekulare Energiewanderung Und Fluoreszenz. *Annalen Der Physik*, **2**(6):55–75 (1948).

Gast K.; Modler A.J. Studying Protein Folding and Aggregation by Laser Light Scattering. Wiley-VCH Verlag GmbH & Co. KGaA (2008). doi:10.1002/9783527610754.sa07.

Ha T.; Enderle T.; Ogletree D.; Chemla D.; Selvin P.; Weiss S. Probing the interaction between two single molecules: Fluorescence resonance energy transfer between a single donor and a single acceptor. *Proceedings of the National Academy of Sciences of the United States of America*, **93**(13):6264–6268 (1996).

Hoffmann A.; Kane A.; Nettels D.; Hertzog D.E.; Baumgartel P.; Lengefeld J.; Reichardt G.; Horsley D.A.; Seckler R.; Bakajin O.; Schuler B. Mapping protein collapse with single-molecule fluorescence and kinetic synchrotron radiation circular dichroism spectroscopy. *Proceedings of the National Academy of Sciences of the United States of America*, **104**(1):105–110 (2007).

Jaenicke R. Stability and folding of ultrastable proteins: eye lens crystallins and enzymes from thermophiles. *The FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, **10**(1):84–92 (1996).

Jia Y.; Talaga D.; Lau W.; Lu H.; DeGrado W.; Hochstrasser R. Folding dynamics of single GCN4 peptides by fluorescence resonant energy transfer confocal microscopy. *Chemical Physics*, **247**(1):69–83 (1999).

Klein-Seetharaman J.; Oikawa M.; Grimshaw S.B.; Wirmer J.; Duchardt E.; Ueda T.; Imoto T.; Smith L.J.; Dobson C.M.; Schwalbe H. Long-range interactions within a nonnative protein. *Science*, **295**(5560):1719–1722 (2002).

Levinthal C. How to fold graciously. *Mössbauer Spectroscopy in Biological Systems Proceedings*, **67**(41):22–24 (1969).

Millett I.S.; Doniach S.; Plaxco K.W. Toward a taxonomy of the denatured state: Small angle scattering studies of unfolded proteins. In G.D. Rose, editor, *Advances in Protein Chemistry*, pages 241–262. Academic Press (2002). doi:doi:10.1016/S0065-3233(02)62009-1.

Moerner W.E. A Dozen Years of Single-Molecule Spectroscopy in Physics, Chemistry, and Biophysics. *The Journal of Physical Chemistry B*, **106**(5):910–927 (2002).

Möglich A.; Joder K.; Kiefhaber T. End-to-end distance distributions and intrachain diffusion constants in unfolded polypeptide chains indicate intramolecular hydrogen bond formation. *Proceedings of the National Academy of Sciences of the United States of America*, **103**(33):12394–12399 (2006).

Nettels D.; Müller-Späth S.; Küster F.; Hofmann H.; Haenni D.; Rüegger S.; Reymond L.; Hoffmann A.; Kubelka J.; Heinz B.; Gast K.; Best R.B.; Schuler B. Single-molecule spectroscopy of the temperature-induced collapse of unfolded proteins. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(49):20740–20745 (2009).

Neuman K.C.; Block S.M. Optical trapping. *The Review of scientific instruments*, **75**(9):2787–2809 (2004).

Onuchic J.N.; Luthey-Schulten Z.; Wolynes P.G. Theory of protein folding: the energy landscape perspective. *Annual Review of Physical Chemistry*, **48**:545–600 (1997).

Pande V.S.; AYu G.; Tanaka T.; Rokhsar D.S. Pathways for protein folding: is a new view needed? *Current Opinion in Structural Biology*, **8**(1):68–79 (1998).

Plotkin S.S.; Onuchic J.N. Understanding protein folding with energy landscape theory. Part I: Basic concepts. *Quarterly reviews of biophysics*, **35**(2):111–167 (2002).

Rief M.; Gautel M.; Oesterhelt F.; Fernandez J.; Gaub H. Reversible unfolding of individual titin immunoglobulin domains by AFM. *Science*, **276**(5315):1109–1112 (1997).

Rischel C.; Poulsen F. Modification of a Specific Tyrosine Enables Tracing of the End-to-End Distance During Apomyoglobin Folding. *FEBS letters*, **374**(1):105–109 (1995).

Sanchez I.C. Phase Transition Behavior of the Isolated Polymer Chain. *Macromolecules*, **12**(5):980–988 (1979).

Scheraga H.A.; Khalili M.; Liwo A. Protein-folding dynamics: overview of molecular simulation techniques. *Annual Review of Physical Chemistry*, **58**:57–83 (2007).

Schuler B. Single-molecule fluorescence spectroscopy of protein folding. *ChemPhysChem*, **6**(7):1206–1220 (2005).

Schuler B. Application of single molecule Förster resonance energy transfer to protein folding. *Methods in molecular biology (Clifton, N.J.)*, **350**:115–138 (2007).

Schuler B.; Eaton W.A. Protein folding studied by single-molecule FRET. *Current Opinion in Structural Biology*, **18**(1):16–26 (2008).

Schuler B.; Lipman E.A.; Eaton W.A. Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature*, **419**(6908):743–747 (2002).

Schuler B.; Lipman E.A.; Steinbach P.J.; Kumke M.; Eaton W.A. Polyproline and the "spectroscopic ruler" revisited with single-molecule fluorescence. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(8):2754–2759 (2005).

Sherman E.; Itkin A.; Kuttner Y.Y.; Rhoades E.; Amir D.; Haas E.; Haran G. Using fluorescence correlation spectroscopy to study conformational changes in denatured proteins. *Biophysical Journal*, **94**(12):4819–4827 (2008).

Shortle D.; Ackerman M.S. Persistence of native-like topology in a denatured protein in 8 M urea. *Science*, **293**(5529):487–489 (2001).

Stryer L.; Haugland R.P. Energy transfer: a spectroscopic ruler. *Proceedings of the National Academy of Sciences of the United States of America*, **58**(2):719–726 (1967).

Thirumalai D.; Lorimer G. Chaperonin-mediated protein folding. *Annual Review of Biophysics and Biomolecular Structure*, **30**:245–269 (2001).

Wilkins D.K.; Grimshaw S.B.; Receveur V.; Dobson C.M.; Jones J.A.; Smith L.J. Hydrodynamic radii of native and denatured proteins measured by pulse field gradient NMR techniques. *Biochemistry*, **38**(50):16424–16431 (1999).

Wu Y.; Kondrashkina E.; Kayatekin C.; Matthews C.R.; Bilsel O. Microsecond acquisition of heterogeneous structure in the folding of a TIM barrel protein. *Proceedings of the National Academy of Sciences of the United States of America*, **105**(36):13367–13372 (2008).

Yang W.Y.; Larios E.; Gruebele M. On the extended beta-conformation propensity of polypeptides at high temperature. *Journal of the American Chemical Society*, **125**(52):16220–16227 (2003).

Ziv G.; Haran G. Protein Folding, Protein Collapse, and Tanford's Transfer Model: Lessons from Single-Molecule FRET. *Journal of the American Chemical Society*, **131**(8):2942–2947 (2009).

Chapter 2

Temperature-induced collapse of unfolded peptide chains

2.1 Introduction

As the initial state of protein folding the study of the conformational states that comprise the unfolded ensemble is of great interest. Recently, with the discovery of the importance of natively unfolded proteins, this has been expanded to include biological function within the unfolded state. In contrast to the relatively static structure of native proteins, unfolded chains undergo major transitions and rearrangements with changing solution conditions. Trying to understand these transitions is essential for gaining further insight into protein folding and function. Single-molecule techniques are especially well suited for the study of heterogenous ensembles, as they allow the observation of the unfolded state in presence of even a dominating folded state.

One feature of unfolded proteins is their collapse with decreasing concentration of denaturants, such as guanidinium chloride or urea. This aspect of unfolded proteins is reminiscent of simple homopolymers, which decrease in dimension as the solvent moves from a 'good' to a 'bad' solvent. However, homopolymers expand as the temperature of the solvent is increased (Sun *et al.*, 1980), in contrast to the chemically more complex chains of peptides. Most importantly, proteinogenic amino acids have varying degrees of hydrophobicity and charge, and the burial of hydrophobic side chains in native-like solvents is one of the major driving forces of protein folding. Also specific interactions lead to the formation of the native structure under folding conditions, and might also contribute to the unfolded state dimensions, e.g. in the form of residual secondary structure.

In this study we combine the ability of single-molecule FRET spectroscopy to study the unfolded state of proteins, even in the presence of the native state and regardless of solvent

conditions with a systematic variation of chain composition to elucidate how temperature induced polypeptide chain collapse differs from homopolymer behavior.

2.2 Materials and Methods

2.2.1 Preparation and labeling of λ repressor

The primary amino acid sequence of the pseudo wild-type λ repressor, as found in the PDB structure '1LMB', was used as a basis to generate an *E. coli* codon usage-optimized construct, which was expressed in the Novagen vector pET-47b(+), containing a cleavable hexahistidine tag. Cysteine residues were placed at positions 6 and 84, respectively (λ P6C I84C), to allow for the covalent dye attachment. Sequence synthesis and subcloning were carried out by Celtek Bioscience LLC. The protein was mainly expressed in inclusion bodies of *E.coli* BL21 cells in LB medium with 1 mM Kanamycin at 37 °C over night. Harvested cells were resuspended in 100mM Tris buffer, 1 mM EDTA pH 8 and subjected to disruption in the presence of Complete Protease Inhibitor Cocktail and benzonase for DNA/RNA digestion. The resulting suspension was mixed with 0.5 volumes of 60 mM EDTA, 6% (w/v) Triton X100, 1.5 M NaCl pH 8 and vigorously stirred for at least 2 h at 4 °C. The inclusion bodies were pelleted by centrifugation, and washed 2-4 times with the same buffer. The pellet was then dissolved in 20 mM Tris buffer, 0.5 M NaCl, 6 M guanidine hydrochloride (GdmCl), 20 mM Imidazole, 2 mM β-mercaptoethanol at pH 8 (IMAC-buffer). This solution was directly applied to a 5ml GE Healthcare HisTrap HP Nickel-IMAC column. An imidazole gradient from 20 to 300mM was used to elute the protein. Protein-containing fractions were pooled and concentrated via ultrafiltration (Amincon Centricon 3000 Da MWCO). To cleave off the histidine tag, the protein was rapidly diluted into the appropriate buffer for cleavage with HRV 3C (50mM Tris buffer, 150mM NaCl, pH 7.5) and digested for at least 2 h at room temperature. After ultrafiltration and buffer exchange to IMAC-buffer, uncleaved protein and the protease where removed by another HisTrap column. The flow-through fractions were collected and concentrated by ultrafiltration again. Just before the labeling, the cysteine residues were reduced by a large excess of DTT (150 mM). Monomeric reduced protein was separated by size exclusion chromatography using a Superdex 75 10/300 GL column (GE Healthcare Biosciences) in 50 mM sodium phosphate buffer, 6 M GdmCl, 150 mM NaCl. The correct molecular weight of the unlabeled protein was confirmed by electronspray ionization mass spectrometry. The concentration was determined by UV absorption spectroscopy. The fluorescent dyes, Molecular Probes Alexa Fluor 488 C5-malemide (donor) and Alexa Fluor 594 C5-malemide (acceptor), were each dissolved in DMSO to a concentration of 20 µg/µl and treated with ultra sonic for at least 20 min. The protein was reacted with the donor in a 1:0.6 stoichiometric ratio for

2 h at room temperature under nitrogen atmosphere. At that point also the acceptor dye was added at a ratio of protein to dye of 1:10 and the reaction incubated over night. Protein and unreacted dye were separated by size exclusion chromatography analogous to the previous step. Correct labeling, existence of the three possible permutations unlabeled, singly labeled and doubly labeled, was confirmed by ESI-MS.

2.2.2 Circular dichroism measurements

In preparation of the CD measurements on unlabeled λ repressor, size exclusion chromatography was performed to ensure that the protein was reduced and monomeric (in 50 mM sodium phosphate buffer, 3 M GdmCl, 100 mM NaCl, 1 mM TCEP, pH 7; protein was injected with 50 mM DTT). The concentration was determined to be 82 µM. The high concentration of GdmCl ensured fully unfolded protein and minimal aggregation. CD measurements were limited to wavelengths between 210 and 250 nm due to the absorbance of GdmCl. The temperature was varied from 5 to 95 °C. Afterwards a comparison spectrum at 5 °C was measured on the same sample to check for baseline changes due to aggregation and peptide bond cleavage. The same experiment was performed with buffer to account for temperature induced changes in the background. Similar experiments with constant measurements of the ellipticity at 222 nm, while heating from 5 to 95 °C, showed similar results (data not shown).

2.2.3 Single molecule measurements

Samples were stored in 8 M GdmCl, and freshly and rapidly diluted in 50 mM sodium phosphate buffer with 0.001% Tween 20 and 150 mM β-mercaptoethanol to a protein concentration range of around 15 pM. The actual concentration was adjusted in a way that we obtained approximately 10.000 to 40.000 photon bursts in a 30 min interval using a burst identification method based on an increase of the photon frequency (Eggeling *et al.*, 2001) with a 50 photon cutoff and a maximum time between successive photons of $<$100 µs in a burst. The single molecule measurements and the calibration of the custom-built temperature-controlled sample holder were carried out as described in Nettels *et al.* 2009. The laser power for the measurements was set to 120 µW at 485 nm. To minimize any changes in the solution conditions over the measurement, the protein solution of 100 µl was overlaid with 40 µl mineral oil (Sigma). Each temperature point was measured for 30 min, starting with the lowest temperature on the same sample. At temperatures above 60 °C, samples were frequently exchanged by aliquots from the same stock solution.

2.2.4 Analysis of single molecule data

Single molecule data were recorded with the Symphotime software (PicoQuant GmbH) and analyzed using the custom FRETica package for Mathematica. The Förster distance was corrected for temperature as in Nettels *et al.* 2009. To obtain the mean transfer efficiencies of the unfolded populations represented in the transfer efficiency histograms, we fitted them with normal distributions. Native state populations and donor only distributions were fitted with log-normal distributions. The position, width and asymmetry of the latter were fixed to the values fitted at the lowest temperature. In cases were populations were overlapping, the contribution of the unfolded state to the overall histogram was enriched using the RASP analysis by Hoffmann *et al.* 2011. The time at which the 'same molecule' probability drops below 60% was calculated from every measurement. Histograms from bursts of recurring molecules in this time span and in an appropriate transfer efficiency range, where most of the folded and donor only population is excluded, were generated. These histograms then were fitted as described above.

2.2.5 Polymer theory

The calculation of the overall interaction energies from transfer efficiencies was performed as in Müller-Späth *et al.* 2010 and Ziv and Haran 2009. Briefly, we used the mean-field theory derived by Sanchez (Sanchez, 1979) to obtain an expression for the end-to-end distance distribution $P(r)$ of a self-avoiding chain. The theory gives an expression for the probability density function of the radius of gyration R_g, $P(R_g)$, as a function the intra-chain interaction energy ϵ and the radius of gyration of the protein at the Θ-state, R_{g_Θ}. The Θ-state is defined as the state of a polymer in which attractive and repulsive forces within the chain and with the solvent balance out and the polymer obeys the length scaling of an ideal chain.

$$P_{Sanchez}(R_g) = P_0(r_g)\exp(Nq(\phi_s,\epsilon)) \tag{2.1}$$

This equation consists of two components: first, a distribution of radii of gyration r_g for an ideal chain $P_0(r_g)$, which is approximated by the Flory-Fisk distribution (Flory and Fisk, 1966)

$$P_0(r_g) \propto r_g^6 \exp(-{}^7\!/_2\, r_g^2/\langle R_{g_\Theta}^2\rangle) \tag{2.2}$$

where R_{g_Θ} is assumed to be $R_{g_\Theta} = 0.22\,\text{nm}\, N_{\text{bonds}}^{1/2}$ (Hofmann, personal communication) with $N_{\text{bonds}} = N + 9$ being the number of peptide bonds between the dyes, taking the fluorophore linker into account (McCarney *et al.*, 2005), and a Boltzmann factor whose value depends on the excess free energy per monomer with respect to the ideal chain

$$q(\phi_s,\epsilon) = {}^1\!/_2\epsilon_{\phi_s} - (1-\phi_s)\ln(1-\phi_s)/\phi_s \tag{2.3}$$

where ϵ (in k_BT) is the mean interaction energy between amino acids; ϕ_s is the volume fraction of the chain relative to the most compact state as in $\phi_s = R_{g,N}^3/R_g^3$. The most compact state is $R_{g,N} = (N+1)v/(4\pi/3)^{1/3}$, where v is the volume of an amino acid (0.13 nm^3).

In order to relate the distribution of radii of gyration obtained from the Sanchez theory with the end-to-end-distance distribution that is used to describe the single-molecule data, we use a conditional probability distribution of the distance between two random points in a sphere of a given radius of gyration (Ziv and Haran, 2009):

$$p(r|r_g) = \frac{1}{\delta \cdot r_g}\left(3\left(\frac{r}{\delta \cdot r_g}\right)^2 - \frac{9}{4}\left(\frac{r}{\delta \cdot r_g}\right)^3 + \frac{3}{16}\left(\frac{r}{\delta \cdot r_g}\right)^5\right) \tag{2.4}$$

with $\delta = \sqrt{5}$, which was determined for the condition that $\langle r^2 \rangle = 6R_g^2$ at the Θ-point. With the Förster equation $E(r) = R_0^6/(R_0^6+r^6)$, where R_0 is the Förster distance, and equations 2.1 and this 2.4, we can describe the mean transfer efficiencies obtained from transfer efficiency histogram peaks:

$$\langle E \rangle = \int_0^L \left(\frac{R_0^6}{R_0^6 + r^6}\right) \int_{R_{G,N}}^{L/2} p(r|r_g) P_{Sanchez}(r_g)\, \mathrm{d}r_g dr \tag{2.5}$$

where L is the contour length of the chain (segment). We solve this equation for ϵ to obtain the interaction energies in the chain.

2.3 Results

2.3.1 The unfolded state of λ repressor collapses with increasing temperature, but re-expands

Much like the small all-β-sheet protein Csp*Tm* and the intrinsically disordered protein prothymosin α , the unfolded state of the helix bundle protein λ repressor collapses with increasing temperature. The low stability of this variant enabled us to observe both the unfolded and the native state at the same time in the virtual absence of denaturants, as shown in figure 2.1 A. This eliminates the need for extrapolating the unfolded state dimension to native-like conditions from a series of denaturant dependent experiments. We also can follow the shift from the folded to the unfolded population due to unfolding with increasing temperature. As the unfolded state becomes more compact, native and unfolded state distributions overlap. Thus, the need for a better way to separate them arises. We employ the RASP method to select bursts originating from the unfolded state (figure 2.1 B). Combining this method with a kinetic model also allowed us to rule out fast interconversion dynamics between the unfolded state ensemble and the native state up to a temperature of 319K (data not shown), which otherwise would have contributed to the shape of the transfer efficiency histograms. However,

the differences in the fitted mean transfer efficiencies of the unfolded state are small between histograms comprised of all or selected bursts only. In contrast to all other reported protein chains, the λ repressor shows a collapse to a clear minimum, followed by a re-expansion of the chain. Unfortunately, the sample longevity is severely decreased at high temperatures hindering us from exploring the dependency at even higher temperatures.

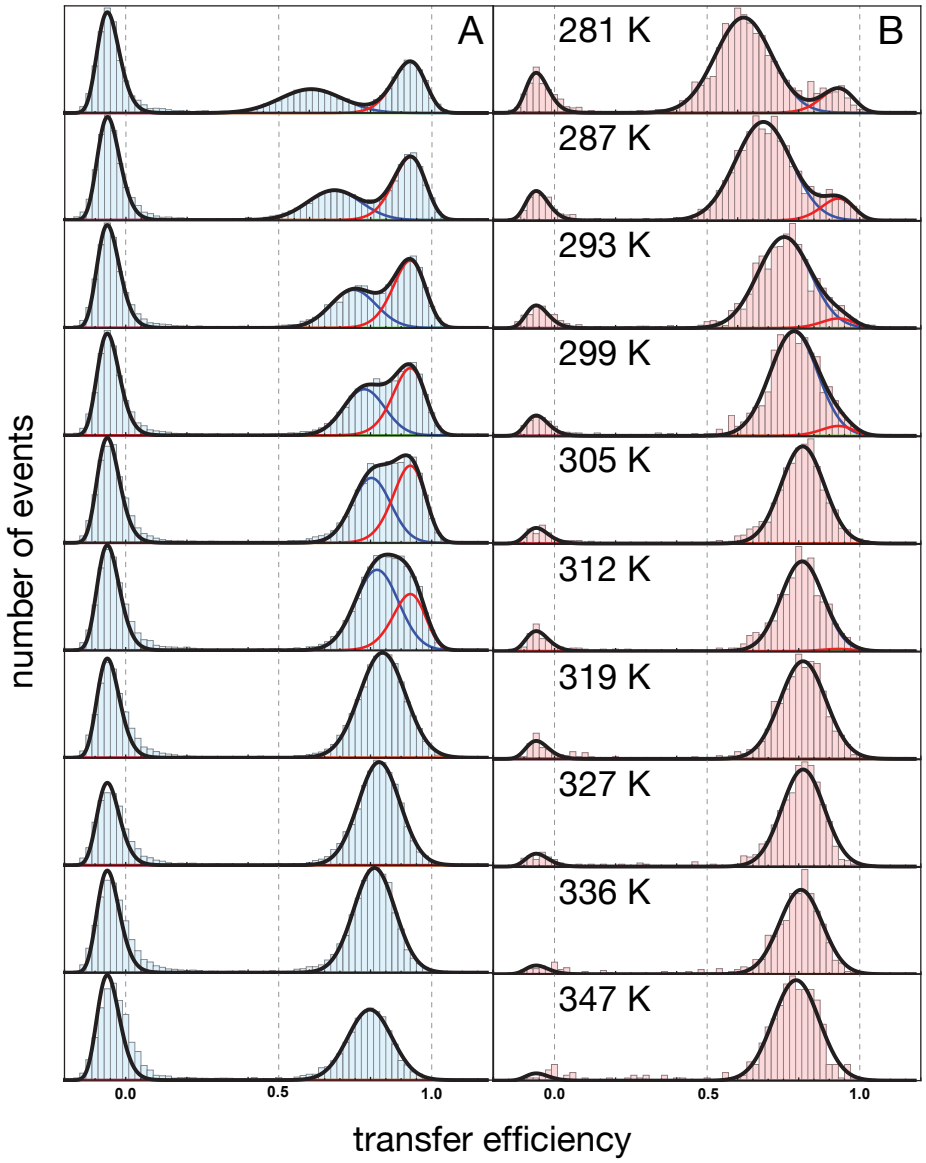

Figure 2.1: Typical data set of equilibrium transfer efficiency histograms at increasing temperatures, here showing λ-repressor R82C P6C. A Set of full histograms fitted with 3 populations B Histograms with an enriched contribution of the unfolded state, by RASP with a same molecule probability threshold of 60%.

2.3.2 Potential contribution of secondary structure to the collapse of lambda P6C I84C might be smaller than anticipated

It has been suggested (Yang *et al.*, 2003) that in several proteins with different secondary structure content and folds, extended beta-structures emerge at high temperatures. Yang *et al.* investigated circular dichroism spectra and traces at a wavelength of 222 nm of λ repressor in high concentrations of GdmCl. They concluded that changes in the CD signal at 222 nm of 2–3 mdeg μM^{-1} cm^{-1} in 6 M GdmCl are indicative of extended structure propensity. We performed similar experiments (see figure 2.2) with our variant of λ repressor, and only see changes of 1 mdeg μM^{-1} cm^{-1} at 222nm in 3 M GdmCl, which is close to the noise in the spectra. Unfortunately, circular dichroism measurements in the absence of denaturants have to deal with a significant population of the the folded state, which obviously would contribute to the CD-signal. The lack of denaturants also leads to aggregation at higher temperatures. In 3 M GdmCl, the sample yields a spectrum comparable to the one at the start of the experiment after being cooled down again.

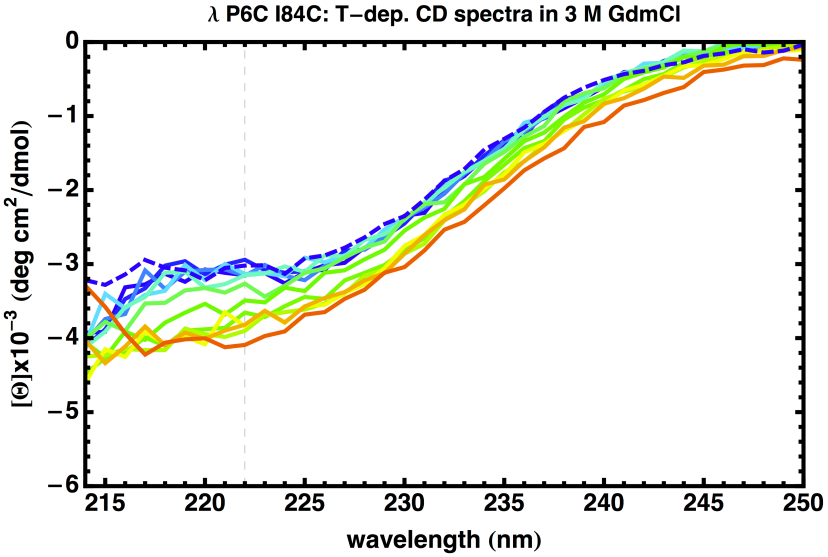

Figure 2.2: Molar Ellipticity spectra from a heating experiment of chemically unfolded λ P6C I84C at temperatures from 5 °C to 95 °C. The dashed line shows the spectrum derived after cooling the sample down again to 5 °C

2.3.3 Temperature collapse is a general property of unfolded protein chains

In order to determine if the temperature-induced collapse is a generic property of polypeptide chains, and to relate its characteristics to the chain composition, we expanded the study to other proteins: Two variants of the intrinsically disordered protein (IDP) prothymosin α with

different charge contents of the sequence between the two cyteines used for dye attachment; N-terminal domain of HIV integrase, an IDP in which native-like structure can be induced upon zinc binding; and a 34 amino acid fragment of the two-state folder Csp*Tm*, which is not folding component. Each of these polypeptides allows us to study its unfolded state ensemble under native-like conditions. In the absence of denaturants or specific binders, these proteins show no folded state, so we only used RASP histogram purification if the unfolded population overlapped by the proteins lacking an intact acceptor dye. Although all these proteins vary considerably in amino acid composition, and thus in hydrophobicity, as well as in charge density and distribution, they all exhibit a temperature-induced unfolded state collapse as shown in figure 2.3. However, λ repressor remains the sole protein to display a minimum of its radius of gyration in the accessible temperature range, even though the other chains also show a curvature. We convert FRET efficiencies to mean field interaction energies within the chain using the theory of Sanchez (Sanchez, 1979). The resulting free energy is positive for chains with predominantly attractive interactions and negative for repulsive interactions. This is illustrated in figure 2.3, were strongly negatively charged chains substantially differ in terms of energy from more uncharged and hydrophobic ones.

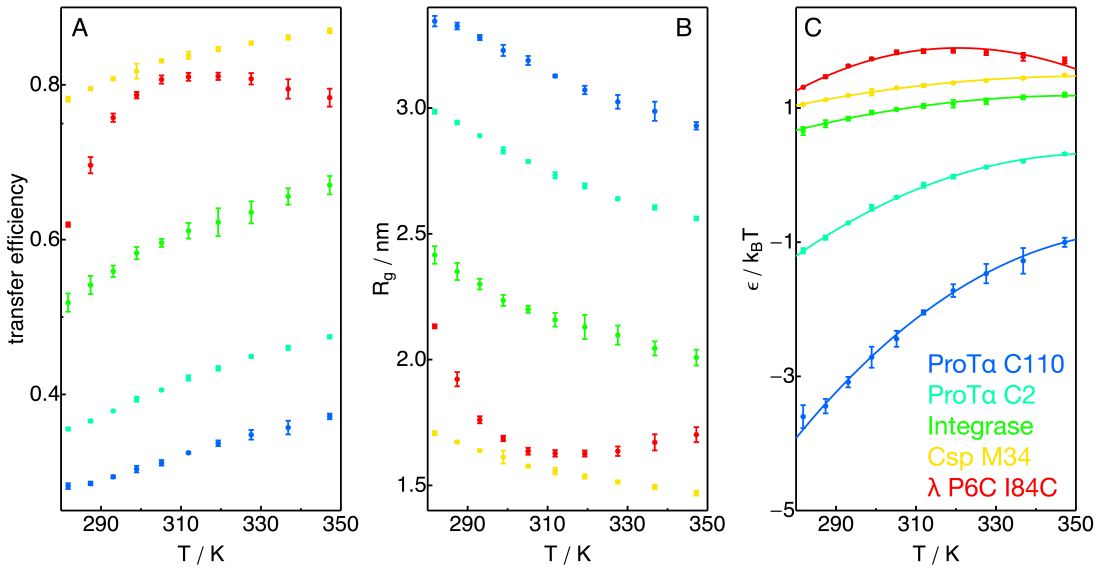

Figure 2.3: Transfer efficiency (**A**), radius of gyration (**B**) and intra-chain energy (**C**) derived by applying the Sanchez-Model vs. temperature. The energies are fitted according to equation 2.6. Data and fits are colored according to the respective protein names.

2.3.4 The change in free energy of chain collapse with temperature shows a complex behavior

The intra-chain energy ϵ we now treat as a change in free energy with temperature as in Chan and Dill 1991 Eq. 40 (Nettels *et al.*, 2009):

$$\Delta G = \Delta H_0 + \Delta C_p \left(T - T_0\right) - T \left(\Delta S_0 - \Delta C_p \log(T/T_0)\right) \tag{2.6}$$

where ΔH_0 and ΔS_0 are the enthalpic and entropic contributions to the collapse process, respectively, at the reference temperature T_0 (298 K) assuming a constant heat capacity ΔC_p. Thus we obtain these contributions, as well as the the heat capacity, of the change in the intra-chain energy in the collapse process as presented in table 2.1. From that also the temperature of maximum free energy (T_M) can be calculated ($T_M = \exp(-(\Delta S_0/\Delta C_p))T_0$). For λ repressor, integrase and Csp M34 the entropic component ($-T \left(\Delta S_0 + \Delta C_p \log(T/T_0)\right)$) of the process is slightly larger, than the enthalpic one ($\Delta H_0 + \Delta C_p \left(T - T_0\right)$). For prothymosin α C110 this relation is reversed and for prothymosin α C2 it changes over the temperature range. The temperature of maximum free energy can only be observed with λ repressor, but is not in the accessible temperature range with the other proteins.

To get insights into the effect of sequence composition on the thermodynamic parameters, we correlate thermodynamic parameters with the mean charge and the mean hydrophobicity represented by the average Kyte-Doolittle score (Kyte and Doolittle, 1982), a widely used scale for the hydrophobicity of peptides, of the chain between the cysteines to which the fluorescent probes are attached (figure 2.4).

Protein	$\Delta H_0/k_BT$	$\Delta S_0/k_B$	$\Delta C_p/k_B$	T_M / K
prothymosin α C110	-19.6	-0.056	0.27	368.5±3.6
prothymosin α C2	-9.8	-0.031	0.18	355.7±1.6
integrase	-2.4	-0.011	0.07	350.8±4.8
Csp M34	-1.4	-0.09	0.05	354.9±3.8
λ repressor	-3.5	-0.017	0.24	320.5±0.8

Table 2.1: Thermodynamic parameters of the temperature collapse at the reference temperature of 298 K. Energies as in figure 2.3 C were fit according to equation 2.6

2.4 Discussion

In the present study we increase our knowledge of the effect of temperature on unfolded proteins. This work expands on the the study published by Nettels *et al.* 2009. In this work the

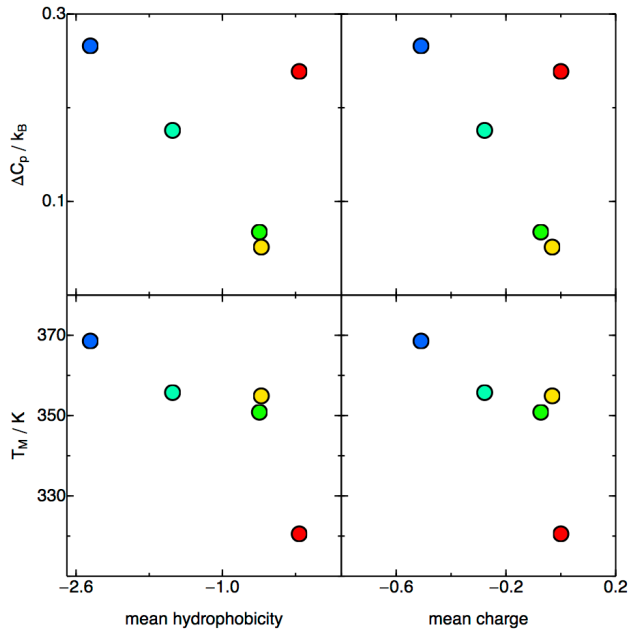

Figure 2.4: Plot of the derived change in heat capacity ΔC_p and the maximum "turnover" temperature T_M vs. either the average Kyte-Doolittle score or mean charge of the the chain between the cysteines. The filled circles are colored as in figure 2.3.

authors found a collapse of the unfolded state of Csp*Tm* with temperature in the range of 280-340 K using single-molecule techniques and dynamic light scattering. Since all Csp*Tm* is folded in the absence of denaturants, measurements were carried out at different concentrations of GdmCl, and their results were extrapolated to physiological buffer. This was not necessary in the present study, due to the fact that all investigated proteins show an unfolded population or are completely unfolded in.

The hydrophobic effect gains in strength until approx. 400 K (Dill, 1990), and is a potential source of an interaction compacting the protein. However, the intrinsically disordered protein prothymosin α, which has a much lower hydrophobicity compared to Csp*Tm*, also exhibited temperature-induced collapse. Csp*Tm* showed some slight changes in the CD signal at 222 nm hinting at some involvement of secondary structure in the collapse. Similarly, a set of molecular dynamics simulations showed a collapse in dimension and an increase in intramolecular hydrogen bonding, but was very dependent on the choice of force field and water model.

We have shown that temperature-induced chain collapse extends over a variety of unfolded polypeptides of different types (IDPs, two state folders, protein fragments), suggesting it to be a general property of protein chains. This was done under single molecule conditions in the absence of denaturants, eliminating signal contribution from the folded state and effects

of binding of GdmCl to the chain. Chain collapse persists independent of whether attractive or repulsive forces, due to electrostatic interactions, dominate. Furthermore we have shown that unfolded λ repressor has a clear minimum in chain dimensions, a behavior that has also been shown for the solvation of small hydrophobic molecules at different temperatures (Dill, 1990; Privalov, 1988), this may indicate that the hydrophobic effect has a maximum also in protein chains, followed by chain re-expansion at elevated temperatures. λ repressor behavior therefore shows characteristics of both hydrophobic hydration of small molecules and homopolymer expansion.

Furthermore we show on purely unfolded λ repressor that residual secondary structure is, if at all, only a small contributor to temperature collapse. However, it is important to note that this experiment does not rule out the existence of secondary structure in the unfolded state or that it contributes to collapse, but merely shows that a significant component of the change in chain dimension is driven by a different mechanism. While the exact relation between degree of collapse, or rather the interaction energy to maintain the collapsed conformation, and hydrophobicity or charge can not be derived from the data, it is clear that they are correlated, however not in a simple linear fashion.

However, it is hard to predict and measure how polymers as complex as proteins are solvated. Attempts to reproduce the polymer collapse and correct chain dimensions in molecular dynamics simulations have been somewhat successful, but are very dependent on the employed water model and its combination with a protein force field (Best and Mittal, 2010). This is not surprising since a complete description calls for a water model that can replicate pure water properties and hydrogen bond networks near different surfaces accompanied by an optimized force field that can describe not only native, but also unfolded chains. To further complicate things, polypeptides are not simple homopolymers, but are composed of residues of varying hydrophobicity and charge in different patterns. This leads to solvation of hydrophobic cavities of different sizes within the water network, which are separated in sequence. It has been suggested that this leads to different regimes of hydration (Huang and Chandler, 2000; Chandler, 2005). This might explain the special behavior of λ repressor. It is not only more hydrophobic than the other investigated proteins, but its hydrophobic residues are also fairly large and somewhat enriched in the C-terminal half.

This might promote the formation of unstructured clusters of hydrophobic residues in the unfolded state. It already has been suggested from NMR experiments (Klein-Seetharaman *et al.*, 2002) that such clusters exist. Currently Dr. Robert Best and Dr. Jeetain Mittal are performing explicit solvent molecular dynamics simulation, to look at possible cluster formation of hydrophobic residues and the water structure around them. Further experiments with random peptides of different average hydrophobicity and residue pattering could elucidate this

phenomena as well.

Bibliography

Best R.B.; Mittal J. Protein simulations with an optimized water model: cooperative helix formation and temperature-induced unfolded state collapse. *The Journal of Physical Chemistry B*, **114**(46):14916–14923 (2010).

Chan H.S.; Dill K.A. Polymer principles in protein structure and stability. *Annual review of biophysics and biophysical chemistry*, **20**:447–490 (1991).

Chandler D. Interfaces and the driving force of hydrophobic assembly. *Nature*, **437**(7059):640–647 (2005).

Dill K.A. Dominant forces in protein folding. *Biochemistry*, **29**(31):7133–7155 (1990).

Eggeling C.; Berger S.; Brand L.; Fries J.R.; Schaffer J.; Volkmer A.; Seidel C.A. Data registration and selective single-molecule analysis using multi-parameter fluorescence detection. *Journal of biotechnology*, **86**(3):163–180 (2001).

Flory P.; Fisk S. Effect of Volume Exclusion on Dimensions of Polymer Chains. *The Journal of Chemical Physics*, **44**(6):2243–& (1966).

Hoffmann A.; Nettels D.; Clark J.; Borgia A.; Radford S.E.; Clarke J.; Schuler B. Quantifying heterogeneity and conformational dynamics from single molecule FRET of diffusing molecules: recurrence analysis of single particles (RASP). *Phys. Chem. Chem. Phys.*, **13**(5):1857–1871 (2011).

Huang D.M.; Chandler D. Temperature and length scale dependence of hydrophobic effects and their possible implications for protein folding. *Proceedings of the National Academy of Sciences of the United States of America*, **97**(15):8324–8327 (2000).

Klein-Seetharaman J.; Oikawa M.; Grimshaw S.B.; Wirmer J.; Duchardt E.; Ueda T.; Imoto T.; Smith L.J.; Dobson C.M.; Schwalbe H. Long-range interactions within a nonnative protein. *Science*, **295**(5560):1719–1722 (2002).

Kyte J.; Doolittle R.F. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, **157**(1):105–132 (1982).

McCarney E.R.; Werner J.H.; Bernstein S.L.; Ruczinski I.; Makarov D.E.; Goodwin P.M.; Plaxco K.W. Site-specific dimensions across a highly denatured protein; a single molecule study. *Journal of Molecular Biology*, **352**(3):672–682 (2005).

Müller-Späth S.; Soranno A.; Hirschfeld V.; Hofmann H.; Rüegger S.; Reymond L.; Nettels D.; Schuler B. Charge interactions can dominate the dimensions of intrinsically disordered proteins. *Proceedings of the National Academy of Sciences of the United States of America*, **107**(33):14609–14614 (2010).

Nettels D.; Müller-Späth S.; Küster F.; Hofmann H.; Haenni D.; Rüegger S.; Reymond L.; Hoffmann A.; Kubelka J.; Heinz B.; Gast K.; Best R.B.; Schuler B. Single-molecule spectroscopy of the temperature-induced collapse of unfolded proteins. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(49):20740–20745 (2009).

Privalov P. Stability of protein structure and hydrophobic interaction. *Adv Protein Chem* (1988).

Sanchez I.C. Phase Transition Behavior of the Isolated Polymer Chain. *Macromolecules*, **12**(5):980–988 (1979).

Sun S.; Nishio I.; Swislow G.; Tanaka T. The coil-globule transition: Radius of gyration of polystyrene in cyclohexane. *The Journal of Chemical Physics*, **73**(12):5971–5975 (1980).

Yang W.Y.; Larios E.; Gruebele M. On the extended beta-conformation propensity of polypeptides at high temperature. *Journal of the American Chemical Society*, **125**(52):16220–16227 (2003).

Ziv G.; Haran G. Protein Folding, Protein Collapse, and Tanford's Transfer Model: Lessons from Single-Molecule FRET. *Journal of the American Chemical Society*, **131**(8):2942–2947 (2009).

Chapter 3

Folding dynamics of λ-repressor

3.1 Introduction

3.1.1 Theory and basic considerations

The aim of this part of the thesis was to watch proteins fold – one molecule at a time and resolve the transition paths. As previously mentioned, this would, in theory, yield information about all pathways a protein can take from the unfolded to the folded state and their probabilities and kinetics. Are there certain successive events with respect to secondary and tertiary structure during the acquisition of the native fold? Does a single pathway dominate the reaction or are parallel routes possible, and if so, how do they relate to one another quantitatively? Whilst a multidimensional exploration of all possible conformations would be the desired goal, both simulation-based and experimental approaches have faced considerable hurdles. In theory, all spatiotemporal information would be accessible *via* all-atom molecular dynamics simulations, but capturing the pathway ensemble is computationally expensive due to the long time scales required for folding. Long continuous all-atom explicit-water simulations above 1 μs have not been feasible until recently (Lindorff-Larsen *et al.*, 2011), resulting in extremely rare observations of folding events under native conditions. A possibility to improve the statistics is the reduction of the representation of the protein and solvent structures, e.g. implicit water model or residue bead models. However, the nature of this coarse graining of the simulation can lead to a bias of the results. Recently, some advances have been made in reconstructing full pathways from many short atomistic simulations (Noé *et al.*, 2009). These indicated that protein folding is a highly parallel process along several pathways. On the experimental side, single molecule FRET can access the time scales of folding more easily, but is limited to probing the pair-wise[1] interactions of donor and acceptor labeled residues. Care-

[1] This recently has been expanded to multiple intramolecular distances by three (Hohng *et al.*, 2004) or even four (Lee *et al.*, 2010) color single-molecule FRET.

fully selected, the distance between the residues can provide a suitable observable to monitor the folding process. With a very high photon rate, one could resolve individual folding events in single molecule trajectories and dissect them into successive events: the dynamics in the unfolded state, ascending the energy barrier, traversing the transition state and reaching the native state.

A free energy barrier between the unfolded and native state has been observed for the majority of proteins studied experimentally (Thirumalai *et al.*, 2010); it is the characteristics of this barrier, i.e. height and roughness, that determine the kinetics of the folding reaction. This leads to two- or multistate kinetics. A notable exception to this is the downhill folding regime, which will be discussed later.

In the context of single-molecule FRET experiments we are posed with the challenge to reconcile the timescales of folding and barrier transition with the data sampling rate imposed by the FRET process and the characteristics of confocal detection. In single molecule FRET experiments we work in the range of excitation saturation of the donor dye. Although the timescale of the fluorescence lifetime of the dyes is in the range of a few nanoseconds, we only detect photons approximately every 10 µs. This is due to non-emitting states such as the triplet state, long-lived ionization states or quenched states, in which at least one of the dyes reside most of the time during a photon-burst (Ha and Tinnefeld, 2012). On the other hand the barrier crossing process is in the low microsecond range (Kubelka *et al.*, 2004). Therefore, on average, barrier transitions will not be resolved. However, the photon emission is a stochastic process. Therefore it is possible that on some events the dyes will not enter any non-emitting states and one will be able to collect many photons during the diffusional crossing from the same molecule. This needs to coincide with the barrier transition happening. Since it is relatively easy to accumulate high number of recorded photon bursts it is reasonable to assume that it is possible to observe such events.

3.1.2 Folding rates and transition path times

Several timescales are associated with the folding process, but not all are easily accessible experimentally. In recent years it has become possible to determine the reconfiguration time in the unfolded state (Nettels *et al.*, 2007, 2008). Another important parameter is the folding rate, which is the rate at which the protein reacts from the unfolded ensemble to the native state. Folding rates can be described by Kramers' theory for unimolecular reactions (Kramers, 1940; Hänggi *et al.*, 1990). It approximates the reaction as the diffusion on a free energy surface along a reaction coordinate. Despite the multidimensional nature of the energy landscape, Kramers' equation has been used to describe the rate of transition from the

unfolded well to the native state in protein folding (Zwanzig, 1997):

$$k_f = \frac{\omega_{min}\omega_{max}D_{max}}{2\pi k_B T} \exp\left(-\Delta G^{\ddagger}/k_B T\right) = k_0 \exp\left(-\Delta G^{\ddagger}/k_B T\right) \tag{3.1}$$

where k_f is the folding rate; ω_{min} and ω_{max} describe the curvatures at the unfolded well and at the barrier top, respectively; D_{max} is the diffusion coefficient at the transition state; $\Delta G^{\ddagger}$ is the height of the barrier; T is the temperature and k_B is the Boltzmann constant. The above parameters constitute the Kramer pre-exponential factor k_0. Figure 3.1 illustrates the approximation of the folding reaction by the Kramers equation.

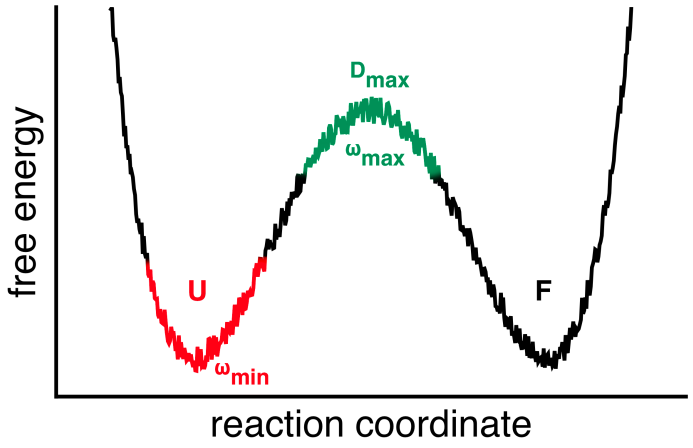

Figure 3.1: Illustration of the Kramers equation Protein folding is understood as a diffusive process along a reaction coordinate. The rate of the process is determined by the height of the barrier and the prefactor, which is influenced by the shape of this landscape.

Another interesting time scale is the time it takes the protein to cross the energy barrier, the so called transition path time. The folding rate does not directly reveal this time, apart from the fact that the rate of folding can not be faster than the inverse transition path time. Obviously the transition path time can also not be faster than the time it takes two residues in a chain to form a contact.

The average duration of transition paths crossings $\langle t_{TP} \rangle$ along a reaction coordinate x can be calculated analytically (Attila Szabo, personal communication cited in Hummer 2004; Chung *et al.* 2009)

$$\langle t_{TP} \rangle = \int_{x_0}^{x_1} e^{-G(x)/k_B T} \phi_U(x) \phi_F(x)\, \mathrm{d}x \int_{x_0}^{x_1} e^{-G(x')/k_B T}\, \mathrm{d}x' \,/ D \tag{3.2}$$

where x_0 and x_1 correspond to the reaction coordinates in the unfolded (U) and folded (F) state, respectively; $G(x)$ is the free energy, ϕ_U and ϕ_F are the fractions of trajectories starting from x_0 and x_1 that reach U and F first, respectively; D is the diffusion coefficient, which

is assumed to be independent of the position along the reaction coordinate. For barriers $>$ $2k_BT$ the above equation yields:

$$\langle t_{TP} \rangle \approx \frac{\ln[2e^{\gamma}\Delta G^{\ddagger}/k_BT]}{D(\omega^{\ddagger})^2/k_BT} = \frac{\ln[2e^{\gamma}\ln(k_0/k_f)]}{2\pi k_0} \tag{3.3}$$

where $\gamma = 0.577$ (Euler's constant), $(\omega^{\ddagger})^2$ characterizes the curvature of the barrier, and k_0 is the pre-exponential factor in equation 3.1. It has been estimated that transition path times lie in the range of 0.6 to 62 µs for the protein GB1 (Chung *et al.*, 2009). Very recently the transition path time was determined for FiP35 WW domain to be approx. 2 µs (Chung *et al.*, 2012). These times are much faster than the measured inverse folding rate coefficients. Also they differ from each other much less then the folding rates (5x versus 10000x).

3.1.3 Downhill folding and typical model proteins

A special and very interesting type of proteins is the class of downhill folders (Bryngelson *et al.*, 1995). In their folding reaction, any major energy barrier is absent (Garcia-Mira *et al.*, 2002) or has been removed by means of protein engineering (Sabelko *et al.*, 1999). Downhill behavior is expected to occur under conditions of extreme native bias, while for the same protein at elevated temperatures and higher denaturant concentrations, the folding mechanism might tend to switch to two-state behavior. However, some studies show the lack of folding barriers in some proteins under a wider range of solvent conditions, leading to so called one-state folding (Li *et al.*, 2009; Liu *et al.*, 2012). Downhill folding leads to the possibility that subpopulations of partly folded conformations, belonging to a transition state ensemble, can be significantly occupied at any point on the reaction coordinate, not just in the unfolded and folded wells. As a consequence, these proteins could be especially beneficial for single molecule experiments (Eaton, 1999), as they will spend more time in the transition region of the energy landscape, during the rather limited observation time. In the downhill folding scenario, in contrast to a barrier-limited approach, one can assume that $\Delta G^{\ddagger}$ (see equation 3.1) is negligible, and as a result of these factors, downhill folding permits the exciting opportunity to elucidate the otherwise obscure pre-exponential factor of Kramer's theory – the protein folding speed limit. Downhill folding has so far been suggested for the miniprotein BBL (Garcia-Mira *et al.*, 2002), dimeric yeast transcription factor GCN4 (Meisner and Sosnick, 2004), villin headpiece subdomain (Lei *et al.*, 2008; Godoy-Ruiz *et al.*, 2008), yeast phosphoglycerate kinase, human ubiquitin (Sabelko *et al.*, 1999), WW domain (Liu *et al.*, 2008), and and some variants of λ-repressor (Ma and Gruebele, 2005; Dumont *et al.*, 2006; Liu and Gruebele, 2007), as will be discussed below.

3.1.4 Strategies for experimental determination of folding rates and transition path times

With respect to FRET-based experiments, with dyes covalently attached to the polypeptide, two possibilities of observation have found application in the folding community. In a lot of studies, proteins are immobilized by different means (on a surface or in a host matrix) and observed until the bleaching of any of the chromophores. In this case, folding times can be derived by the waiting time distributions in different E-states. Although this procedure seems rather simple at first, it bears in itself the difficulty of avoiding additional conformational constraints imposed by the surface and artifacts due to interaction of the dyes with the matrix. Studies have tried to mitigate these effects by encapsulating the proteins in surface tethered lipid vesicles (Rhoades *et al.*, 2003, 2004) or by elaborate data analysis to better understand how the trajectories are influenced by dye photophysics and immobilization (Chung *et al.*, 2009, 2010). Also, while immobilized proteins in principle could be observed virtually forever, the maximum time of a trajectory is limited by the bleaching of the dyes or their transition into long-lived non-fluorescent states. The bleaching rate is heavily dependent on the excitation intensity, which in turn determines photon rate and the potential time resolution of the experiment.

An alternative to the approach utilizing protein immobilization is the recording of short trajectories of freely diffusing molecules. This avoids any disturbing surface effects on protein and dyes. However, this method prevents the collection of long continuous information from a single molecule, making the determination of folding rates not as straightforward. Some possible methods will be discussed in section 3.4. However, since one is limited in the observation time to the diffusion time through the confocal volume, one can increase the laser power to a value so that bleaching will not artificially shorten the majority of the trajectories, but the photon rate, and therefore time resolution is enhanced. Another constraint on the design of experiments on freely diffusing proteins is the necessity that in order to study transition path times, the barrier transitions have to occur during the diffusion through the confocal volume. Therefore the diffusion time and the folding rate have to be in the same regime in order to achieve sufficient statistics on the barrier transitions. Ideally one wants to avoid much higher folding rates, which would result in several transitions per burst, in order to simplify data analysis. The number of transitions per time is also related to the relative populations of the unfolded and folded state. One would expect the maximum frequency at the solvent conditions of the ensemble transition midpoint where $k_f = k_u$. Under these conditions the proteins will continuously transition between the unfolded and folded state, instead of residing in one state. The midpoint is however the condition of the slowest relaxation rate constant of the system (minimum of $k_{obs} = k_f + k_u$). Therefore it is a major challenge in

the study of transition path times to find a model protein system that satisfies the demand of having (un)folding rates at the unfolding midpoint matching the diffusion time as closely at possible.

3.1.5 λ repressor

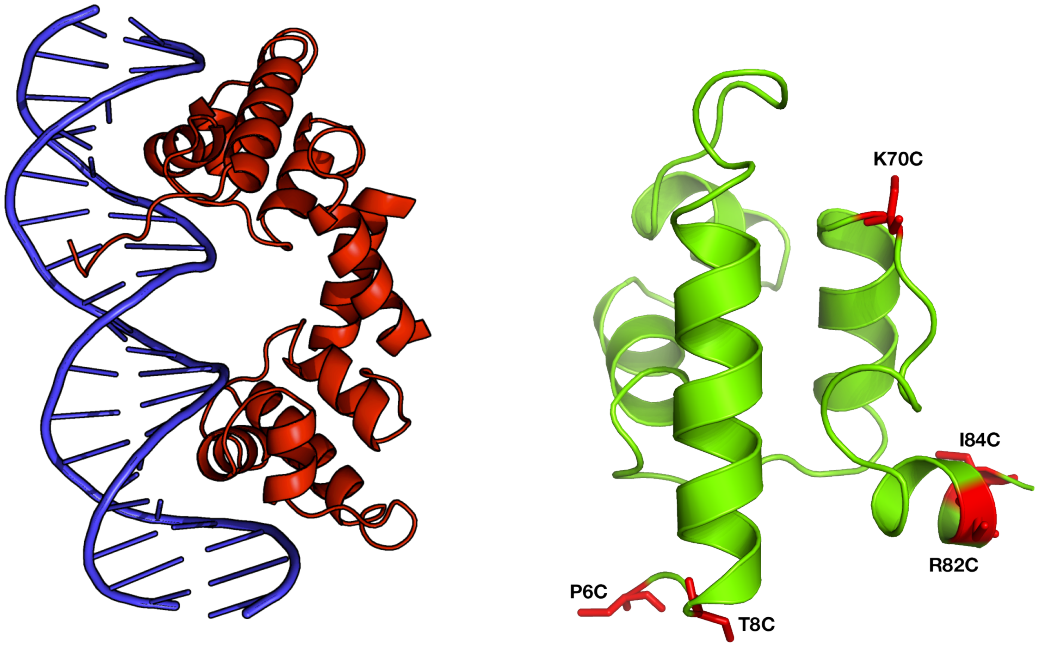

Figure 3.2: Cartoon models of the N-terminal domain of the λ repressor (PDB ID '1LMB', Beamer and Pabo 1992). On the left the full N-terminal domain is shown, binding its operator DNA sequence as a protein dimer. On the right side the $\lambda_{6\text{-}85}$ fragment is shown, which is the variant that has been studied predominately in the context of protein folding. The N-terminal tail and the part of the fifth helix, which facilitates dimerization, have been truncated. The residues shown in red have been mutated to cysteines in the different variants in order to attach the fluorescence dyes.

λ repressor has been studied extensively in the past two decades (Burton *et al.*, 1997; Yang and Gruebele, 2003). In the bacteriophage λ, the *c*I protein suppresses the switch from lysogenic growth to the lytic cycle by binding to its chromosome (Spiegelman *et al.*, 1972; Meyer *et al.*, 1975; Ackers *et al.*, 1982). The *c*I repressor is a homodimer, with each monomer consisting of two subunits and a total of 237 residues (Stayrook *et al.*, 2008). From this protein, the "λ repressor" peptide has been derived, comprising the N-terminal DNA binding domain as seen in figure 3.2. The wild-type and several variants of this protein have been studied in the folding community. Specifically, the peptide 6-85, as solved in the PDB structure '1LMB' (Beamer and Pabo, 1992), with small alterations at the termini, has primarily been used as

a pseudo wild-type, with other variants adapted from it. The folded λ repressor is predominately alpha-helical with several loops. Of the five alpha helices, the first four contribute to the hydrophobic core packing to varying degrees (Lim *et al.*, 1994), whilst the fifth helix plays a major role in dimerization of the full protein. This helix is usually truncated at residue 85 to prevent unwanted dimerization. The folding kinetics of numerous variants of this polypeptide was studied extensively with various methods. Two research groups have been especially successful in their endeavors to reveal several aspects of λ repressor folding and properties. The first is the group of Terrence Oas, Duke University. Employing primarily NMR- and CD-based methods, they were able to show that the pseudo wild-type $\lambda_{6\text{-}85}$ exhibited two state folding behavior (Huang and Oas, 1995a). The method of NMR lineshape analysis allows the determination of both folding and unfolding rates even in equilibrium experiments. Using the ^{1}H aromatic NMR spectra they were able to quantify rates in the presence of varying concentrations of urea, which were extrapolated to submillisecond folding times under native conditions (Huang and Oas, 1995b). Subsequently, they proposed that λ repressor folds *via* a compact transition state (Burton *et al.*, 1996). They later moved on to a more complete description of the folding landscape of the wild type and the thermostable variant G46A/G48A using a series of alanine to glycine mutations (Burton *et al.*, 1997). They further investigated a tryptophan-containing variant additionally using stopped-flow methods (Ghaemmaghami *et al.*, 1998), specific buried hydrogen bonds (Myers and Oas, 1999) and mimicked the unfolded state under native conditions using methionine oxidation (Chugha *et al.*, 2006; Chugha and Oas, 2007). A second major contributor to biophysical λ repressor research is the group of Martin Gruebele, University of Illinois at Urbana-Champaign, who made great use of laser-induced temperature jump experiments (Gruebele *et al.*, 1998). They use changes in the fluorescence lifetime of an introduced tryptophan residue upon refolding to obtain kinetic information. In their initial studies (Yang and Gruebele, 2003, 2004), they observed different phases in the refolding reaction, which could be better fitted with double exponential decays compared to single exponential ones. This did not occur in all variants, but only in the very fast folding ones (λ D14A, Q33Y). The interpretation for this effect was that through mutation and choice of solvent conditions (temperature, viscogens), the major barrier in the protein folding reaction was reduced to a degree that the roughness of the energy landscape in the transition state became an apparent and finally dominant factor in the rate of reaching the native state. The determined time scale is approx. 2 μs, which is slower than the reconfiguration time of small proteins in the unfolded state (ns, Nettels *et al.* 2007) and faster than typical folding rates of small proteins that follow a two state kinetic scheme (10s of μs) or multi-domain proteins (seconds and above). In later studies, the group expanded their repertoire of variants and suggested several folding mechanisms from two-state to downhill

for them (Ma and Gruebele, 2005; Liu and Gruebele, 2007).

3.2 Methods and Materials

3.2.1 Sample Preparation

Molecular biology and protein expression

The plasmid for the first variant of λ repressor was generated by site-directed mutagenesis based on a construct kindly provided by Terrance Oas, Duke University. The wild-type sequence of λ repressor is devoid of cysteine residues. In order to facilitate labeling with Alexa dyes, we introduced two cysteine residues, one at position 8 (T8C) and one at postion 82 (R82C). The reasoning behind this placement was that the Oas group had found this pair to be suitable in terms of folding rates at the unfolding midpoint during trial ensemble experiments (Oas, personal communication), albeit in combination with a variant that still possessed the first five wild-type residues. The coding sequence was cloned into the pAED4 expression vector (Doering, 1992). A Stratagene QuikChange site-directed mutagenesis kit was used following the manufacturer protocols to introduce the cysteine residues, with oligonucleotides purchased from by Microsynth. Correct modification was confirmed by sequencing (Micosynth). The coding sequences for the variants T8C/K70C and P6C/I84C where chemically synthesized and subcloned into an empty Novagen pET-47b(+) by Celtek Genes, USA (all sequences can be found in the appendix). This vector has an N-terminal hexahistidine tag followed by a human rhinovirus (HRV) 3C protease cleavage site, which is directly adjacent to the sequence of λ repressor and closing with several stop codons.

All protein expression was started from freshly transformed *E.coli* BL21 (DE3) cells. Single colonies were picked from LB agar plates with the appropriate antibiotic and transferred to a custom medium[2] or LB medium. Cell growth was carried out in 1 l batches in 5 l flasks in a shaker at 110 rpm and 37 °C. Induction was started by addition of IPTG to a final concentration of 1 mM, when the OD at 500 nm reached 0.8. The T8C/R82C and T8C/K70C variants were allowed to grow for additional 3 hours, while it was beneficial to express P6C/I84C over night. The cells were harvested by centrifugation for 20 minutes at 5100 g.

Protein purification

Given the different expression vectors, two strategies for protein purification were developed. T8C/R82C was purified by exploiting physicochemical differences between λ repressor and

[2]pAED4: carbenicillin for media and plates; custom growth medium: 20 g/l Tryptone, 10 g/l yeast extract, 80 mM NaCl, 0.4% (w/v) glucose, 0.36x M9 salts mix; pET-47b(+): kanamycin standard LB medium and plates

endogenous *E. coli* proteins whereas those proteins expressed in pET-47b(+) exploited the presence of the poly-histidine tag for immobilized metal ion affinity chromatography (IMAC). An optimized protocol is presented here:

T8C/R82C

Cells were resuspended in lysis buffer (50 mM Tris-Cl, pH 8.0, 10 mM EDTA, 20 mM DTT) and disrupted with a French press at 4 °C in the presence of protease inhibitors TAME, Benzamidine, PMSF and/or Roche Complete. Nucleic acids were digested by addition of 5 U Benzonase per 1 ml of suspension at 4 °C. This variant was primarily found in the soluble fraction, therefore remaining cell debris was removed by centrifugation for 1 hour at 18,000 g. The supernatant was loaded onto a custom column of the weak anion exchanger DEAE-Sephacel (in 10 mM Tris-Cl, pH 8.0, 2 mM $CaCl_2$, 0.1 mM EDTA, 50 mM NaCl, 15 mM methionine). Somewhat counter-intuitively, T8C/R82C, with a calculated pI of 6.2-6.6, did not bind quantitatively, but instead could be found in the flow-through of the column, while many unwanted proteins were retained. Freezing of the flow-through fractions in liquid nitrogen (or storage at 4 °C over night) led to substantial precipitation of larger proteins, but no precipitation of T8C/R82C. This procedure was used as an additional purification step. Further precipitation of unwanted proteins was achieved by adding 4 M ammonium sulfate to a final concentration of 2 M. The suspension was spun down at 18,000 g for approx. 30 min to sediment the unwanted precipitate. After centrifugation, the supernatant was applied to a hydrophobic interaction column (GeHealthcare HiPrep 16/10 Butyl FF) and eluted in a gradient of ammonium sulfate from 2 M to 0 M in 20 mM Tris-acetate pH 6.2. Fractions containing λ repressor were identified by SDS-PAGE, pooled, concentrated *via* ultrafiltration (Amicon 50 ml cell, 3000 Da MWCO) and further purified by size exclusion chromatography (Superdex 75, in 10 mM sodium phosphate buffer (NaP), 150 mM NaCl, pH 7) in multiple runs. This procedure also removed unwanted aggregated protein. Correct molecular mass was confirmed *via* ESI-MS and Edman-degradation, showing that the N-terminal methionine is cleaved off *in vivo*.

T8C/K70C and P6C/I84C

These variants were purified using the N-terminal hexahistidine tag. Expression lead to the formation of inclusion bodies. Although some protein could also be found in the fermentation supernatant, only the inclusion body pellets were further processed. Omitting the soluble fraction led to an improvement of the purity of the protein of interest in the first few steps of the protocol, when compared to T8C/R82C. Harvested cells were resuspended in 100 mM Tris buffer, 1 mM EDTA pH 8 (wash buffer) and subjected to disruption in the presence of

Roche Complete inhibitor mix and benzonase. Cell debris was kept in solution by adding 0.5 vol of 60 mM EDTA, 6% (w/v) Triton X100, 1.5 M NaCl pH 8 and was vigorously stirred for at least 2 h at 4 °C. The inclusion bodies were pelleted by centrifugation and washed two to four times with wash buffer. The pellet was then dissolved in 20 mM Tris, 0.5 M NaCl, 6 M GdmCl, 20 mM Imidazole, 2 mM β-mercaptoethanol pH 8 (IMAC buffer). This solution was directly applied to a 5ml HisTrap HP Nickel-IMAC column (GE Healthcare). An imidazole gradient from 20 to 300 mM was used to elute the protein. Protein-containing fractions were pooled and concentrated *via* ultrafiltration (Amincon Centricon 3000 Da MWCO). To cleave the hexahistidine tag, the protein was rapidly diluted into a solution containing *Herpes simplex* (HRV) 3C[3] (50 mM Tris buffer, 150 mM NaCl, pH 7.5) and then digested for at least 2 h at room temperature. Adding 0.5 M of arginine reduced the extent of aggregation in this step, allowed for higher protein concentrations, and did not interfere with the cleavage. After ultrafiltration and buffer exchange to the IMAC buffer, uncleaved protein and the protease where removed by further IMAC, where the flow-through fractions, which contained cleaved protein, were collected and concentrated by ultrafiltration again. Reduction of cysteine residues with a large excess of DTT (150 mM) and a final size exclusion chromatography run in 50 mM NaP, 150 mM NaCl, 6 M GdmCl on a Superdex 75 10/300 GL column were performed just before the labeling reaction. Correct mass of the cleaved product was confirmed by ESI-MS.

Protein labeling

All variants were labeled using cysteine/malemide chemistry (Brinkley, 1992) with the Alexa Fluor (Molecular Probes) dyes Alexa 488 as donor and Alexa 594 as acceptor, respectively. Reactions were carried out at pH 7.0 to ensure maximum selectivity for the protein thiol groups. To prepare for labeling, cysteine residues were reduced with excess TCEP or DTT. In order to remove the reducing agent, ensure the selection of the monomeric protein and to transfer the protein to the labeling buffer, size exclusion chromatography was performed as described above. In cases where it was known that the sample was monomeric, a simpler HiTrap Desalting column could be used to change the buffer. Generally, the protein containing fractions from this chromatography step were collected under argon atmosphere to prevent reoxidation of the cysteine residues. Briefly, the variants T8C/R82C and T8C/K70C were labeled in a stepwise manner. The protein was labeled with one dye, the reaction was quenched and the resulting species purified. The singly labeled species was retained and subsequently labeled with the second dye before the doubly labeled product was repurified.

[3]The activity of the protease was not determined, but digestion of most of the λ repressor was confirmed by SDS-PAGE, therefore establishing working amounts and reaction times for HRV 3C. HRV 3C was expressed and purified using standard protocols for His-tagged proteins (Ricarda Hilf, personal communication).

Specifically, to purify the differently labeled species, the differences in the charges of the dyes were exploited, resulting in different binding affinities to a MonoQ 5/50 GL anion exchanger column (GE Healthcare). Beforehand, to transfer the reaction mixture to the ion exchange buffer (20 mM Tris-HCl, pH 8) and concentrate the sample, Ultrafiltration devices (3000 Da MWCO, Amicon) and HiTrap Desalting columns (GE Healthcare) were used. The protein was applied to the ion exchange column and subsequently eluted in a NaCl gradient from 0 to 600 mM over 40 column volumes. To facilitate the second reaction, dye was added to the reduced protein in a concentration of 20 g/l dye in DMSO and incubated either over night at 4 °C or for 2 h at room temperature. The protein-to-dye stoichiometric ratios were 1:0.6-1:1 for the first and 1:2-1:10 for the second reaction. The reaction was stopped by addition of 150 mM of β-mercaptoethanol. It often has been found to be beneficial, in terms of avoiding aggregation and protein binding to the column, if all steps were carried out under denaturing conditions, e.g. anion exchange with the addition of 6 M urea and size exclusion and ultrafiltration in the presence of 6 M GdmCl. A gel filtration step in 50 mM NaP, 6 M GdmCl, pH 7 was performed to remove any remaining excess dye or non-specific aggregates and prepare the protein for long term storage. Finally the double labeled protein was concentrated by ultrafiltration to a final concentration in the range of 1-5 µM. Correct labeling of singly and double labeled protein was confirmed *via* ESI-MS.

The protocol for the labeling of the P6C/I84C variant was simplified and shortened. Its higher pI lead to poor and quite similar binding among the different labeling permutants to ion exchange columns. Therefore single step labeling was performed. The initial reduction of the protein and the first labeling step were identical to the procedure detailed for the other variants. However instead of a stepwise reaction with subsequent purification of the labeling products, the second dye was added directly to the ongoing reaction of the reduced protein with the first dye. The first dye was allowed to react with the protein alone for 2 h at room temperature under nitrogen atmosphere. At this point the second dye was added and both incubated over night at 4 °C. Protein-to-dye stoichiometric ratio was the same and follow-up purification were performed as in the above stepwise procedure. This procedure was inspired by a more thorough method by Daniel Streich *et al.* (Streich, 2010), that determined reaction constants of the individual cysteine residues from reaction on single-cysteine mutants. Based on these constants, they calculated the optimal time after the start of the first reaction for addition of the second dye, in order that the yield of donor-acceptor-labeled protein is maximized.

3.2.2 Ensemble experiments

Chemical denaturant unfolding transitions

Equilibrium ensemble unfolding series were recorded on a HORIBA Jobin Yvon Fluorolog using FRET-labeled protein. Fluorescence spectra upon donor or acceptor excitation were recorded. Titrations typically were carried out in a buffer of 50 mM NaP, 0.01% Tween 20 at pH 7, with typical protein concentrations of 5-10 nM. Individual measurement points were obtained either by separately preparing solutions at the appropriate concentrations of GdmCl or by subsequently exchanging a protein solution under native conditions with a protein solution containing a high concentration of denaturant. Due to the high folding rate reported for λ repressor, one can safely assume a comparatively rapid equilibration of the signal within a few seconds upon complete mixing of the two solutions. This was confirmed by measuring the same sample several times, which yielded unchanged spectra. Pipetting errors were partly compensated by measuring the refractive index of the final sample and correcting for the total difference in GdmCl concentration. Either FRET-labels or intrinsic tyrosine fluorescence were used. Signals were corrected for the wavelength dependence of the detector sensitivity and fluctuations in the excitation light. Fluorescence intensities at donor (515 nm) and acceptor wavelengths (618 nm) as well as E were fitted to a two-state model (Santoro and Bolen, 1988) with independent baselines for the contribution of the unfolded and the native state to the signal. Considering these baselines, the fluorescence as a function of denaturant could be converted to the fraction of folded protein.

Stopped-flow kinetics

Ensemble kinetic experiments were performed either on a pi*-180 or a SX20 stopped-flow spectrometer (Applied Photophysics). They were set up as refolding experiments, in which FRET-labeled protein in high concentrations of denaturants were diluted with the same buffer containing a lower concentration of denaturant, resulting in a final concentration of denaturant corresponding to the protein ensemble unfolding midpoint, which is the condition of interest for single molecule experiments. To follow the folding reaction, the donor fluorescence intensity was monitored by selecting fluorescence light with a 515nm long-pass filter. Since the efficiency of the detectors supplied with both instruments in the red wavelength regime (over 600nm) is very low, there is negligible influence on the signal due to fluorescence light from the acceptor. Resulting experimental data were fitted with a single exponential decay. Under the assumption that the reaction takes place at the folding midpoint (where folding rate k_{fold} and the unfolding rate k_{unfold} are equal), the relation $k_{obs} = 2k_{fold}$ holds.

To assess the correct function of the stopped-flow instrumentation labeled Csp*Tm* was

used. This was needed, because some variants of λ repressor did not produce a change a signal upon dilution of the denaturant at all. Since the folding rate of λ repressors in theory should approach the limit of the dead time of the stopped flow apparatus, the dead time had to be estimated using a fast chemical reaction. The dead-time of the instrument depends on a variety of factors including the dead volume, the volume of protein pushed through the measurement cell and tubings, the velocity of the piston and correct alignment of the flow circuit, and the ratio of protein to buffer solution. To estimate the dead time of an instrument, one can monitor the fluorescence quenching of N-acetyl-L-tryptophanamide (NATA) by N-bromosuccinimide (NBS) (Shastry *et al.*, 1998). By adding an excess of NBS, this second order reaction can be simplified to a pseudo-first order reaction, and the resulting traces can also be fitted with a single exponential. The difference between t_0 and the time at which the different traces intersect is the dead-time of the instrument. However, the dead time already can be simply estimated to some extent by determining the time where individual repeats on one dataset (combination of NATA/NBS concetration) start to superimpose. The dead time was found to be 1-2 ms.

Circular dichroism

Far UV circular dichroism reports on protein secondary structure, as peptides are optically active compounds that have different absorption coefficients for the left and right polarizations of circularly polarized light. The wavelength region of 160-250 nm is sensitive for the change in secondary structure due to the n→π* and π→π* transitions of the peptide bond. Alpha-helical proteins, such as λ repressor, have a maximum at 192 nm and minima at 207 and 222 nm; unstructured peptides have a minimum at 198 nm. Therefore already the shape of the spectrum yields information about the structure of a protein. To assess if the highly destabilized variants can still fold, we tested if it is possible to shift the equilibrium by stabilizing the native state. Therefore we added the stabilizing cosmotropic salt sodium sulfate to diagnose if this would enhance the native pattern. Measurements were carried out on two different JASCO CD spectrometers, J-715 and J-810. To allow temperature control, Hellma 165-QS cuvettes were used, which were connected to a computer-controlled water bath. Experiments were only conducted on unlabeled protein at different temperatures in native buffer (50 mM NaP, 150 mM NaCl, 1 mM EDTA, pH7) and buffer additionally containing 600 mM sodium sulfate. To reach a sufficient signal-to-noise ratio, but still minimize the influence of aggregation, a relatively low protein concentration of 10 µM was used. Additionally, gelfiltration in the respective measurement buffers was performed immediately before the CD-measurement to remove unwanted salts, protein aggregates, or protein dimers. The sample was injected onto the column in very high concentrations of GdmCl (6-8 M) and reducing

agents (typically 150 mM DTT). Thermal unfolding experiments were attempted but were heavily influenced by massive onset of aggregation at higher temperatures, therefore no clear melting temperature could be determined.

3.2.3 Single molecule experiments

Data collection

Single molecule data were collected on a modified MicroTime 200 confocal microscope (PicoQuant, Berlin). The sample was excited using a continuous wave 485 or 488 nm laser (PicoQuant LDH-D-C-485 and Sapphire 488-100 CDRH, respectively) through an Olympus UplanApo 60/1.20W objective. Fluorescence light was collected through the same objective, and after passing it through a 100 nm pinhole and the major dichroic mirror, used to split excitation and fluorescence light, it was separated into parallel and perpendicular polarized light using a polarizing beam splitter. Each component was then split according to color on either donor or acceptor channel by dichroic mirrors (585DCXR; Chroma Technology), and after passing a last set of filters (ET525/50M and HQ650/100; Chroma Technology), it was focused onto avalanche photodiodes (SPCM-AQR-15; PerkinElmer Optoelectonics). Data were collected in T3 mode, recording the arrival time of each photon using a HydraHarp 400 counting card (PicoQuant, Berlin). Measurements were carried out at a power of 100-160 µW, for times from 30 min to several hours. Samples used in single molecule experiments were stored at -80 °C at protein concentrations of 100 nM in 8 M GdmCl (Pierce). Immediately before the measurement, the sample was diluted stepwise to approximately 15 pM in a final buffer of 50 mM NaP, 150 mM β-mercaptoethanol with 0.001% (v/v) Tween 20.

Generation of single molecule FRET efficiency histograms

Raw data were analyzed using custom software, written by Dr. Daniel Nettles (Group Prof. Ben Schuler). Either the standalone package FRET*ikon* or the backend for Mathematica (Wolfram Research, USA) FRET*ica*. Typically, successive photos separated by less than 100 µs were combined into one burst. Identified bursts were corrected for background, differences in quantum yields, the different efficiencies of the detection channels, cross-talk (or "bleed through"; acceptor emission detected in the donor channel and donor emission detected in the acceptor channel), and direct excitation of the acceptor with a matrix approach (Schuler, 2005)

$$\begin{pmatrix} n_{A,0} \\ n_{D,0} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} n'_A \\ n_D \end{pmatrix} + \begin{pmatrix} b_A \\ b_D \end{pmatrix} \tag{3.4}$$

where $n_{A,0}$ and $n_{D,0}$ are the raw photon counts in the respective channels. The background count rates b_A and b_D can be estimated by determining the photon count rates in the parts of the experiments between the bursts, which equates to blank buffer solution. The corrected values n'_A and n_D are multiplied with the matrix a_{ij} describing the effects listed above. The values of this matrix can be estimated for a single molecule instrument by using solutions of acceptor and donor dye with a concentration ratio equal to their ratio of absorption coefficients at the donor excitation wavelength. Inverting the resulting matrix gives the correction matrix $c_{ij} = a_{ij}^{-1}$, which converts the background-subtracted raw counts to the corrected counts. This procedure can be expanded to handle four-channel detection by using a higher rank matrix. The acceptor counts can now be corrected for direct excitation of the acceptor dye, according to $n_A = n'_A - (n'_A + n'_F) / (1 + \epsilon_D/\epsilon_A)$. ϵ_D and ϵ_A are the respective extinction coefficients at the donor excitation wavelength.

Furthermore, it is possible to identify bursts affected by photobleaching of the acceptor as proposed in Hillger *et al.* 2008. To do so, we define a burst asymmetry $\alpha_{DA} = \bar{t}_D - \bar{t}_A$, where $\bar{t}_{A/D} = 1/n_{A/D} \sum_i t_{A/D,i}$ are the respective average detection times of photons during a burst of duration T with photon arrival times $t_{A/D,1} \ldots t_{A/D,n}$. In the limit of an infinite number of photons α_{DA} is zero if both dyes emit continuously. For finite numbers, one has to consider the resulting shot noise of the burst asymmetry. In the case of no photo bleaching, the detection time of a photon is uniformly distributed with a constant probability density function $p(t_i) = 1/T$, so that $\int_0^T p(t_i)\,\mathrm{d}t = 1$. The expectation value for t_i is still then at the center of the burst as in $\bar{t}_D = \int_0^T t_i p(t_i)\,\mathrm{d}t_i = T/2$, with a variance of $\sigma_i^2 = \int_0^T \left(t_i - \bar{t}_i\right)^2 p(t_i)\,\mathrm{d}t_i = T^2/12$. Now this uncertainty of t_i is propagated to the average value $\Delta\bar{t}_{D,A}^2 = \sum_i \sigma^2/n_{D,A}^2 = T^2/\left(12 n_{D,A}\right)$, and finally to the asymmetry:

$$\sigma_{DA} = \frac{T}{\sqrt{3}} \left(\frac{1}{n_D} + \frac{1}{n_A} \right)^{1/2} \tag{3.5}$$

If the observation of dynamics was not of interest, bursts with $|\alpha_{DA}| \geq 2\sigma_{DA}$ were not included in the analysis. However, the burst asymmetry will be influenced by folding dynamics of the observed protein on the time scale of diffusion/burst time. Unfolding into a low FRET-state would resemble a photobleaching event. Therefore asymmetric bursts were not excluded if observation of fast folding dynamics were the objective of the experiment to prevent biased data analysis. Likewise, bursts with negative asymmetry could be selected, to exclusively analyze putative folding events.

Usually, only burst containing 50 or more photons after correction were retained. For these, the FRET efficiency was calculated according to $E = n'_A/\left(n'_A + n'_D\right)$, and they where were organized into histograms. Very big bursts were often excluded to remove bursts potentially stemming from protein aggregates or rare fluorescent macroscopic particles (dust).

However, these never contributed significantly to histogram shape. Since excitation and detection strategies that eliminate the contribution of protein lacking an active acceptor dye, such as pulsed-interleaved excitation (PIE) or alternating excitation (ALEX), where not applied, the resulting histograms comprise two to three populations. A typical example for a histogram is presented in figure 3.3. The population with the lowest E (centered around zero) can be attributed to proteins lacking an active acceptor dye; either it was never attached during labeling, and the protein is singly or doubly labeled with donor dyes, or because it was photo-bleached during a previous passage through the excitation light. A considerable fraction of background bursts from impurities also fall into this E–range. It is termed "donor-only" and its shape in the E-histogram is usually well described by fitting it to a log-normal distribution. With all investigated proteins, the dyes are on average farther apart in the unfolded state than in the folded state, therefore the population with next highest E is due to the proteins in the unfolded state ensemble. A Gaussian distribution can be used to describe the unfolded state population's shape. Depending on the solution conditions, native proteins form a third population with a relatively high transfer efficiency. It is also fitted using a log-normal distribution.

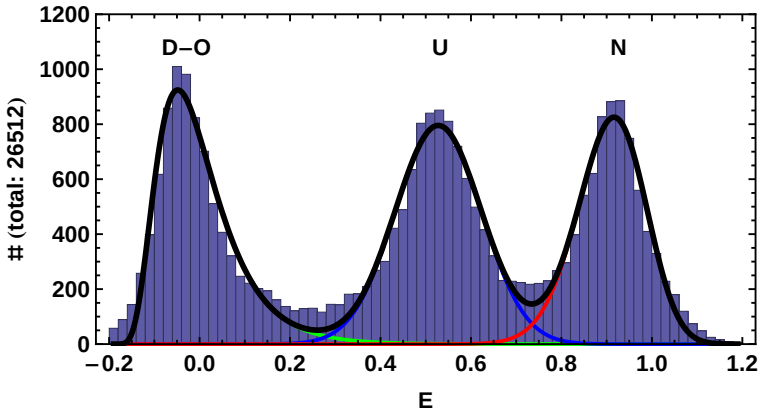

Figure 3.3: Typical example of a single-molecule transfer efficiency histogram. This histogram shows the three populations found in single-molecule experiments. The donor-only (D-O), unfolded (D), and native (N) populations are shown with their fits in green, blue, and red, respectively, and the sum of these fits in black.

3.2.4 Perspective: Microfluidic mixing to determine folding rates

In contrast to equilibrium single-molecule experiments and complementing stopped-flow measurements, we attempted to obtain folding rates at the unfolding midpoint through kinetic measurements in a microfluidic mixing apparatus. The basic principle was a miniaturized continuous flow device, were constant streams of protein and buffer would be mixed at

a ratio that would result in a denaturant concentration corresponding to the unfolding midpoint of the protein. The basis of this technique has been used extensively in the context of single molecule folding (Lipman *et al.*, 2003; Hamadani and Weiss, 2008; Hofmann *et al.*, 2010; Gambin *et al.*, 2011). However, to achieve dead times in the sub-millisecond regime, one needs to move to high flow velocities, making single molecule detection less viable. The folding rate can still be obtained in the ensemble concentration range.

Bengt Wunderlich constructed such a mixer and we performed refolding experiments on λ repressor and Csp*Tm*, and obtained signal traces, which could be fit by a single exponential. However, it turned out out that there was an error in the underlying finite element simulations, which were used to simulate the dilution of the guanidine hydrochloride. The error led to a miscalculation of the diffusion constant of the guanidine hydrochloride, therefore the dilution was not complete at the starting point and the traces likely represent only diffusion of GdmCl out of the sample stream. Another problem is that, to achieve very short deadtimes, the structures on the microfluidic device have to be very small. However, the process (photolithography) in which the templates for the polydimethylsiloxane-based devices are produced could not achieve a high enough fidelity - not correctly representing spatial relations and bending straight lines in structures. Also dust particles clogging channel has been a major problem. In the meantime the photolithography process has been improved, so microfluidic ensemble mixing might be a viable method in a future revision.

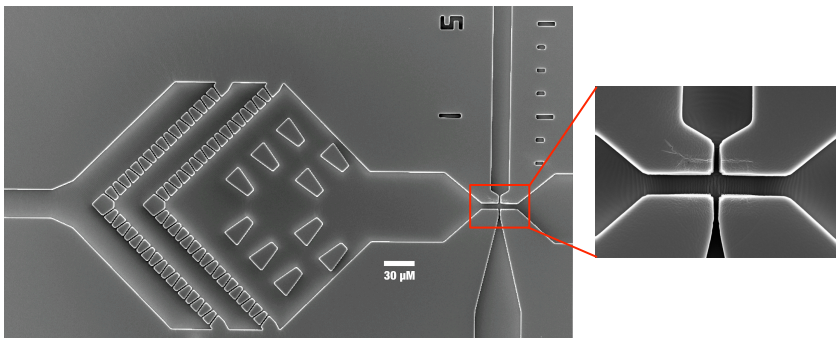

Figure 3.4: Scanning electron microscope image of the microfluidic PDMS based fast ensemble mixer. The mixing region is shown on the right and at higher magnification, with the sample flowing in from the bottom, and the diluting buffer from the sides. The protein kinetics can be observed in the graded region on the top. Every inlet is equipped with a structure shown on the left side. This acts as a filter to prevent remaining dust from entering the mixing region. While the individual channels of this filter are very small, their large number ensures that the backpressure will not increase significantly even if some are blocked. As can be seen at the higher magnification the shape of the mixing region is not entirely perfect, but shows kinks and bends.

3.3 Results

3.3.1 Ensemble and single-molecule unfolding transitions

To characterize the different variants of the labeled λ repressor, protein was unfolded using GdmCl. A series of fluorescence folding experiments at different concentrations of denaturant yields the unfolding midpoint. This point is of particular interest, as these solution conditions maximize the number of potentially observable transition events between the unfolded and folded states. An accurate estimation of the folding midpoint using ensemble techniques is complicated by the instability of λ repressor and its variants. In order to fit the fluorescence signal to a two state model, in ensemble experiments, one needs to quantify the influence of the native and the unfolded state at different concentrations. This procedure requires the native state to be sufficiently populated, in a "pre-transition baseline", before unfolding occurs (equally the unfolded state provides a "post transition baseline"). For λ repressor, no such pre-transition baselines can be measured in standard buffer. The addition of sodium sulfate, which has been shown to preferentially stabilize the native state (Cobos and Radford, 2006) did permit the measurement of a pre-transition baseline, which was then used to fit non-stabilized data using a global analysis method. Assuming a shared pre-transition baseline for different variants has been used successfully in the past, but also includes the inherent assumption that the native states of the different variants behave identically as a function of the denaturant concentration. For the variants T8C/R82C, T8C/K70C and P6C/I84C, unfolding midpoints were determined according to procedures outlined above to be 0.29 M, 0.63 M, and 1.07 M GdmCl, respectively, at room temperature as shown in figure 3.5. Given that single molecule data requires no such baseline analysis, the estimation of the folding midpoints for the variants under investigation obtained from single molecule data was preferred. Especially in the T8C/R82C variant, the signal was influenced by the acceptor dye being quenched in the native state, therefore its midpoint is probably at a higher GdmCl concentration.

Unfolding transitions using GdmCl at room temperature recorded using the single molecule setup are shown in figure 3.6. The P6C/I84C variant exhibits the most common pattern in the histogram, for a two-state folder studied by single-molecule FRET. Three peaks are present in the absence of denaturants. These can be attributed, in order of increasing transfer efficiency, to the donor-only population, the unfolded state ensemble and the native state. In subsequent measurements with increasing concentration of GdmCl, one can observe a decrease of the high-transfer 'native state' peak and a simultaneous increase in the intermediate-transfer 'unfolded state' peak, while the donor-only peak largely remains the same (apart from some problems with impurities as shown in the measurement at 1.66 M). Concurrent with the population shift represented in the histogram the unfolded state ensemble shifts to lower transfer

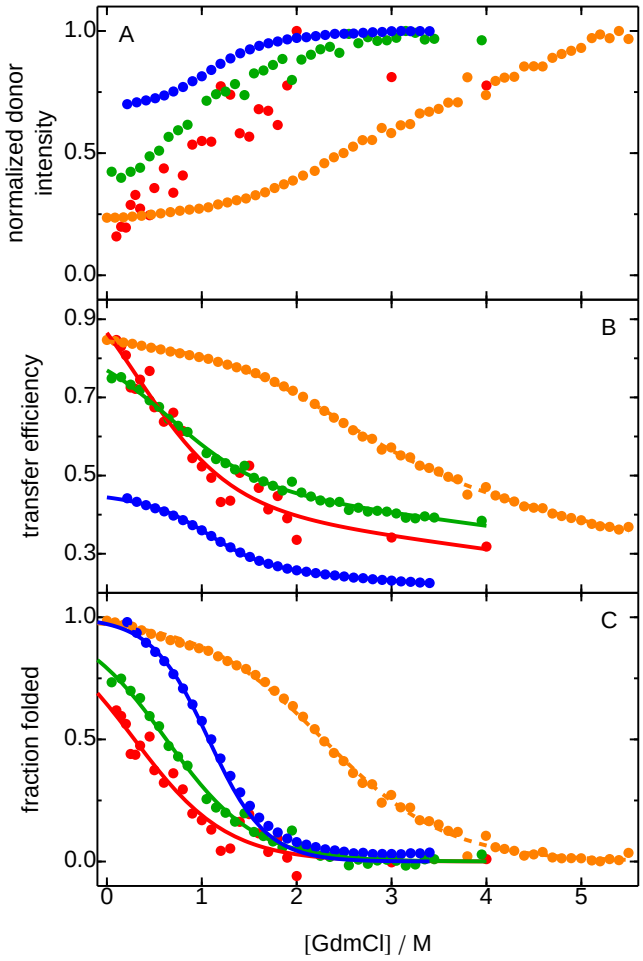

Figure 3.5: Ensemble unfolding transitions of the three λ repressor variants showing the normalized donor intensity - divided by maximum value for clarity (A), transfer efficiency (B) and fraction folding (C) vs. concentration of denaturant (GdmCl). Red: T8C/R82C; Green: T8C/K70C; Blue: P6C/I84C in 50 mM NaP. The orange is T8C/K70C in buffer containing 600mM sodium sulfate. The data sets where fit globally, while only the slope of the pre-transition baseline was shared. However, assigning the pre-transition baseline remains difficult, and therefore the transformation to fraction folded error prone.

efficiencies, corresponding to more compact state of the unfolded protein. This behavior is akin to the findings for Csp*Tm* (Hoffmann *et al.*, 2007). The other variants diverge from this expected pattern in several aspects. The variant T8C/K70C shows, apart from the donor-only peak, only a single peak that shifts from high to low transfer efficiency continuously. However, even from from visual inspection it is clear that the peak is not symmetrical, hinting at different subpopulations which can not be resolved at room temperature. The data presented in figure 3.9 further explored the temperature dimension of this variant, showing that

at low temperatures two clearly distinguishable peaks are present. In accordance with the findings of chapter 3, one can interpret the single peak as a combination of the unfolded and the native population with respective average transfer efficiencies close together enough to complicate the resolution in the histogram. It is worth mentioning that, although a transfer efficiency of the native state of about one was estimated for all variants from the crystal structure, T8C/K70C average transfer efficiency in the native state was considerably lower. The reasons for this remain unclear so far. Finally, the T8C/R82C variant also shows a typical three-state pattern. Despite a step-wise labeling procedure, which greatly reduces the amount of singly and doubly labelled donor-only molecules, the population of native protein at low concentrations of GdmCl is small. As denaturant concentration is increased the native peak largely remains the same, while the unfolded peak increases in relation to the donor-only peak. This hints towards the acceptor dye being quenched in the native state. The detailed reasons for this have not been investigated further. Since the N- and the C-terminus are very close together in the λ repressor, and the dyes are attached near the termini, it is possible that the dyes quench each other. Furthermore, the λ repressor lacks tryptophan residues, which have the potential to quench Alexa dyes (Dominik Hänni, personal communication). There are two tyrosine reissues in the λ repressor sequence, though. Ultimately, the T8C/R82C variant was abandoned, since it further complicates GSML analysis.

3.3.2 Stopped-flow rapid refolding experiments

Successful stopped-flow experiments were only performed on the T8C/R82C variant. Refolding experiments at 0.5, 0.75 and 1 M GdmCl were performed. This yielded traces as shown in figure 3.7. Fitting the data with single exponential decays resulted in folding rate in the regime of approx. 1 ms^{-1}. However, this was very prone to error, introduced by the start time of the data collection, the dead time. In the dead time, the signal does not stem from newly mixed solutions, but the fluctuations of the content in the cuvette (from the ”shot” before). Therefore it needs to be neglected in the trace. Depending on the choice of the cut-off time, the fit yielded different results for the refolding rate as shown in table 3.1. At this point no elaborate dead-time determination was performed, and it was only estimated by inspection of sets of individual traces.

Experiments with the other two variants only yielded constant fluorescence signals upon mixing. However, correct functioning of the stopped flow apparatus was confirmed by performing analogous experiments with labeled Csp*Tm*. Since the relatively slow folding reaction of Csp*Tm* does not report on (sub)millisecond timescales, this procedure was later changed to the quenching reaction of NATA by NBS. To slow down the folding reactions, the temperature was lowered to approx. 4 °C. However, no change was observed. The reasons for this are

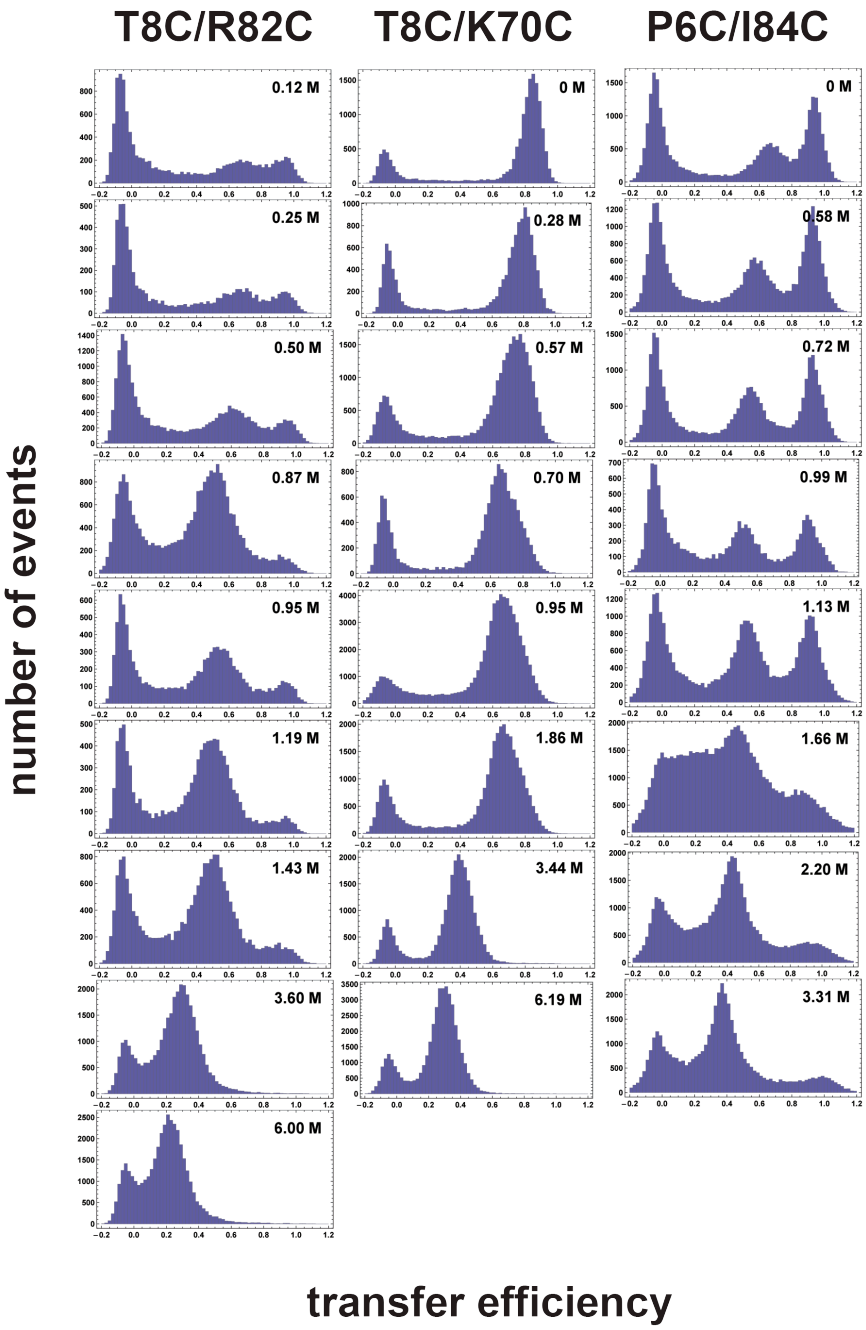

Figure 3.6: Single-molecule GdmCl transitions at room temperature of the three λ repressor variants investigated. Three populations arise. These are in order of the transfer efficiency, the donor-only population, the unfolded state, and the native state. These are clearly visible in the T8C/R82C and P6C/I84C variants. The unfolded state expands with increasing GdmCl concentration, as observed with many proteins. In the T8C/K70C the situation is less clear cut, with only two distinct population apparent. This could be indicative of fast dynamics, However the shape of the high-E distribution changes and further analysis showed that the single population is caused by lack of accuracy of discrimination of unfolded and native state in the histograms.

not clear. To minimize the dead-time, a 1:1 mixing ratio was used. Therefore the difference in GdmCl concentrations between the unfolded protein stock solution and the final mixing solution, at the unfolding midpoint, was only small. This limited the accessible amplitude. In summary, stopped-flow mixing is not an ideal method to probe the fast folding of λ repressor, due to its inherent relatively long dead time. Another ensemble method to study fast kinetics is temperature jumping the solution and following the refolding reaction. Depending on the type of induction of the temperature jump, laser or electric current, different dead-times in the microsecond or even nanosecond range can be achieved.

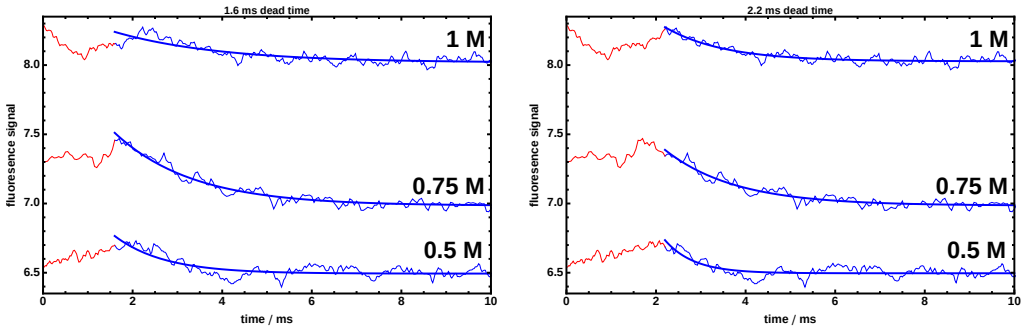

Figure 3.7: Traces from refolding experiments on λ repressor T8C/R82C. Shown are averages from at least 20 repetitions. The signal changes of fluorescence light above 515 nm upon dilution of unfolded protein with native buffer to different final denaturant concentrations were observed. As the protein refolds, donor fluorescence decreases, while the increase of acceptor fluorescence is small up due to poor detection efficiency in the red wavelength regime. The change is fitted to an exponential decay (blue line). Data was fitted only after a certain time point (1.6 or 2.2 ms) to remove the influence of the dead time of the instrument (data shown in red). Since the latter was not determined exactly in this case, this "tailfit" is extremely errorprone. The fitted k_{obs} values for two different dead-times are shown in table 3.1.

3.3.3 Circular dichroism far UV measurements

In order to ensure the capability of T8C/K70C to still be able to assume a folded conformation, spectra with the protein under native conditions and with varying amounts of sodium sulfate were recorded. The spectra exhibit the characteristic shape produced by alpha helical proteins with minima at 222 and 208 nm. Although far-UV CD only reports on secondary structure content and not on tertiary structure and no data of sufficient quality below 200 nm could be recorded, the only modest change in the spectrum (the deepening of the 222 nm minimum) upon addition of sodium sulfate suggests that much of the protein is folded under native conditions. However, it has been shown (Hoffmann *et al.*, 2007) that the unfolded state can contain considerable amounts of secondary structure. Also thermal unfolding was

	refolding rates s^{-1}	
[GdmCl]	1.6 ms dead time	2.2 ms dead time
0.50 M	0.90	1.56
0.75 M	0.58	0.64
1.00 M	0.42	0.76

Table 3.1: Table of determined relaxation rates (k_{obs}) from stopped-flow refolding experiments. The signal changes from refolding experiments at three different final GdmCl concentrations are fitted from time points after 1.6 or 2.5 ms, the assumed dead time of the instrument. At longer (more realistic) dead times, the rate at 0.75 M GdmCl is the slowest, indicating that this denaturant concentration is close to the unfolding midpoint.

performed. However, the CD signal did not reach pre-transition values after cooling of the sample, indicating significant aggregation of the sample at this high protein concentration.

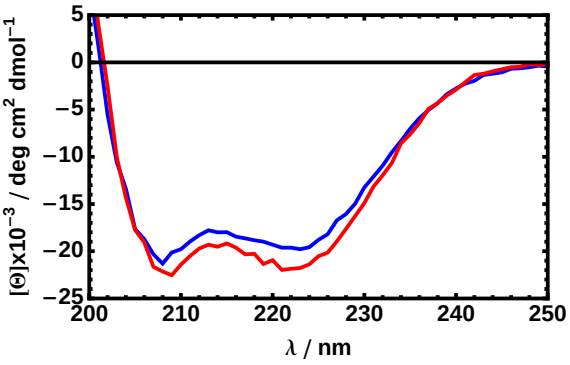

Figure 3.8: Far UV circular dichroism spectra of T8C/K70C. Shown is the corrected normalized molar ellipticity. The protein under native conditions (blue line) shows the characteristic double minima shape of alpha helical proteins. On addition of 600 mM sodium sulfate this is slightly improved (red).

3.4 Analyzing single molecule data in the search for fast dynamics

3.4.1 Introduction and experimental data

Generally, if a system fluctuates between two states, which both give a different signal value that can be observed *via* a measurement method, the relation between the time scale of the signal observation or averaging and the time scale of the state fluctuation will have an influence on the recorded signal. If the state fluctuation is much faster than the integration of

the signal, one will obtain an average signal between the values associated to the different states. On the other hand, the state fluctuation is much slower than the averaging, one will record the two distinct signal values associated to the two different states. Timescales of the state fluctuation between these two extremes will result in broadening of the the signal, somewhere between two distinct signal values and a single average. The extent of this broadening in principle contains the kinetic information of the state fluctuation.

In the context of protein folding, the two states would be the unfolded and the native state, assuming a two state folding protein like λ repressor. The basic principle of determining the state dynamics has been employed with ensemble NMR lineshape analysis (Burton *et al.*, 1997; Myers and Oas, 2001; Arora *et al.*, 2004). However, we can apply the same paradigm to single-molecule FRET measurements. A basic way to analyze single-molecule data from freely diffusing molecules are transfer efficiency histograms. In confocal measurements, the time scale of signal averaging is the diffusion time of the molecules through the observation volume, as all photons of the resulting burst are summed up to determine the transfer efficiency according to equation 1.3. The signal is the transfer efficiency averaged over the burst duration. Therefore in the limits of fast folding dynamics relative to the diffusion time, one would expect a single peak in the transfer efficiency histogram, stemming from proteins that continuously fold back and forth during their diffusion through the confocal volume. In contrast, slow folding proteins would diffuse through the focus in either the folded or the unfolded state and not change during this time, resulting in two peaks in the histogram. Intermediate cases would show broadened peaks (Gopich and Szabo, 2007).

The folding rates can be influenced by solvent conditions: mainly concentration of denaturants - such as GdmCl - and temperature. One would want to stay near the unfolding midpoint to maximize the number of transitions as described before. We measured transfer efficiency histograms at several denaturant concentrations for different variants, as shown in figure 3.6, and looked for broadened peaks. It is apparent that the variant T8C/K70C of λ repressor at all denaturant concentrations only shows a single peak (beside the donor-only peak), though it is clearly not symmetrical. One possible explanation is the presence of fast folding dynamics compared to the timescale of diffusion. One can now change the temperature at which the experiment is carried out and explore the folding behavior in another dimension. Changing the temperature will influence the equilibrium populations, but most importantly will modulate the time scale of the folding dynamics.

In figure 3.9 a series of experiments is shown in which λ repressor T8C/K70C with 0.63M GdmCl, which is approximately the room temperature ensemble unfolding midpoint, was measured at different temperatures. At the lowest temperature, two peaks centered at $E = 0.6$ and $E = 0.8$ are clearly visible. With increasing temperature, these two apparently merge

until only a single peak centered at E=0.7 remains. The observation fits our expectation of fast interconversion between the unfolded and the folded state at higher temperature.

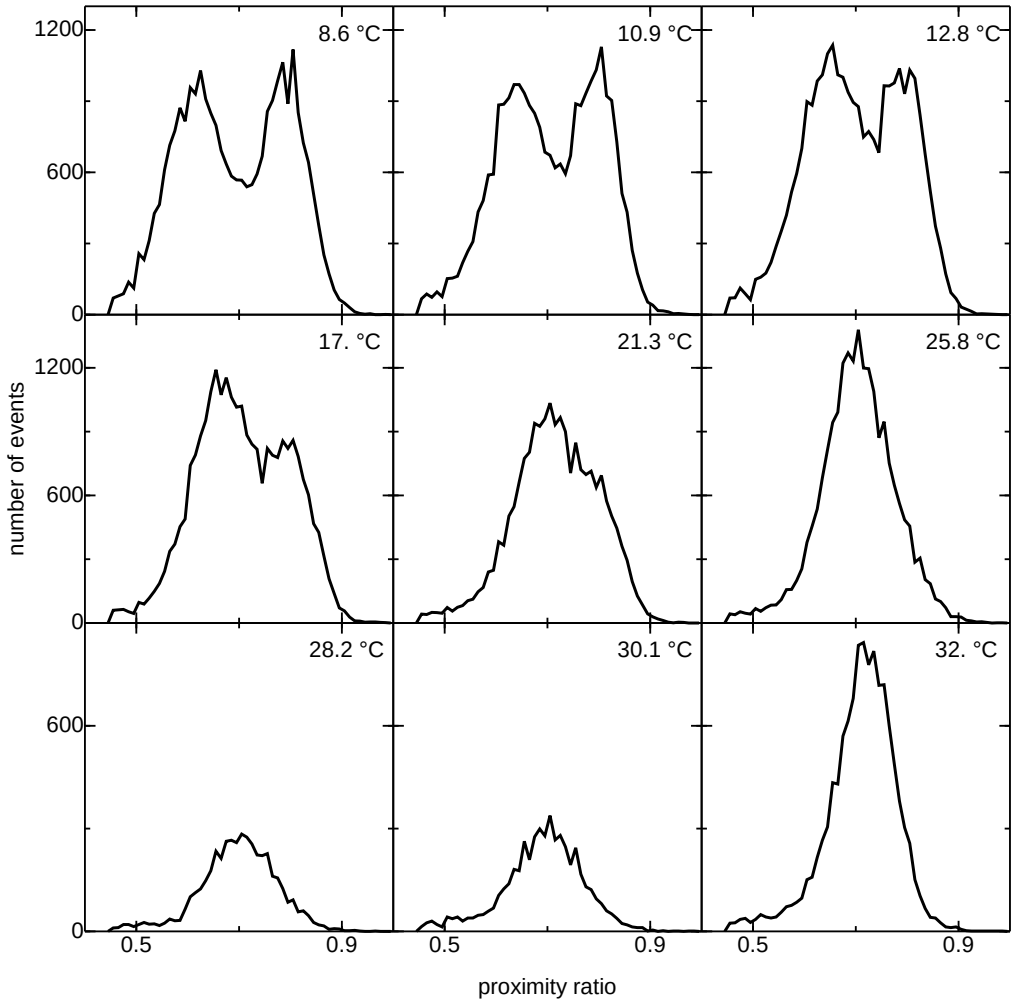

Figure 3.9: Temperature dependent measurement series on λ repressor T8C/K70C. Proximity ratio (uncorrected FRET efficiency) histograms for measurements at different temperatures are plotted. These show the common donor-only peak and two clearly distinguishable peaks at lower temperatures. As the temperature is increased the peeks merge to a single one with a transfer efficiency approximately between the initial two.

3.4.2 Gopich-Szabo maximum likelihood analysis

There are several ways discussed in literature (Gopich and Szabo, 2007; Chung *et al.*, 2009) to determine the interconversion rates from equilibrium single molecule data. We chose a maximum likelihood method that was conceived by Gopich and Szabo (Gopich and Szabo, 2007, 2009) (GSML), for analyzing our data. Compared to the somewhat reduced representation

of a whole photon trajectory in histograms, the Gopich-Szabo maximum likelihood (GSML) analysis is a more rigorous approach, that takes full advantage of the information in the photon color sequence and the interphoton times. This method works under two assumptions about the nature of the photon trajectories. First, that the conformation of the protein is independent of the position in the laser spot. Second, that the total photon count rate of the chromophores is independent of the protein conformation. Consequentially, the interphoton times are related to the translational diffusion of the molecule through the confocal volume. In contrast, the color pattern of the photons stems from the conformational dynamics of the proteins. The GSML method utilizes this color pattern. For a simple two-state folder we assume a kinetic model as in figure 3.10A, in which a folded protein with an average transfer efficiency E_1 unfolds with the rate constant k_u to the unfolded state which has an average transfer efficiency E_2. The folding rate constant is k_f. For a set of these parameters - transfer efficiencies and interconversion rate constants - the likelihood L for the color pattern of a burst with N_{ph} photons and interphoton times $\tau_2 \cdots \tau_{N_{\text{ph}}}$ can be calculated according to:

$$L = \mathbf{1}^T \left[\prod_{k=2}^{N_{\text{ph}}} \left(\mathbf{F}(c_k) e^{\mathbf{K}\tau_k} \right) \right] \mathbf{F}(c_1)\mathbf{p}_{\text{eq}} \tag{3.6}$$

with

$$\mathbf{F}(\text{red}) = \mathbf{E} \qquad \mathbf{F}(\text{green}) = \mathbf{I} - \mathbf{E} \qquad \mathbf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{3.7}$$

$$\mathbf{E} = \begin{pmatrix} E_1 & 0 \\ 0 & E_2 \end{pmatrix} \qquad \mathbf{K} = \begin{pmatrix} -k_u & k_f \\ k_u & -k_f \end{pmatrix} \qquad \mathbf{p}_{\text{eq}} = \begin{pmatrix} \frac{k_f}{k_u + k_f} \\ \frac{k_u}{k_u + k_f} \end{pmatrix} \tag{3.8}$$

where c_k, with $k = 1 \cdots N_{\text{ph}}$, is the 'color' (green/donor or red/acceptor) of the kth photon; the vector $\mathbf{p}_{eq}$ contains the equilibrium populations of the two conformational states; $\mathbf{F}(c_1)$ and $\mathbf{F}(c_k)$ are matrices according to equation 3.7, for the first and subsequent photons; $\mathbf{1}^T$ is the unit row vector, and $\mathbf{I}$ is the unity matrix. An example for this equation, for a short five-photon burst is given in figure 3.10. The logarithm of the likelihood of all N_{bursts} of a measurement can be calculated by:

$$\Delta = \sum_{i=1}^{N_{\text{bursts}}} ln(L_i) \tag{3.9}$$

The model parameters (E_1, E_2, k_f, k_u) are found by maximizing Δ, however we fixed the transfer efficiency of the native state E_1 to 0.8, as it is reasonable to assume that the native state will not change with temperature. The burst size cutoff for the GSML analysis was usually increased to 80-100 photons. Also the length of a typical measurement needed to be increased

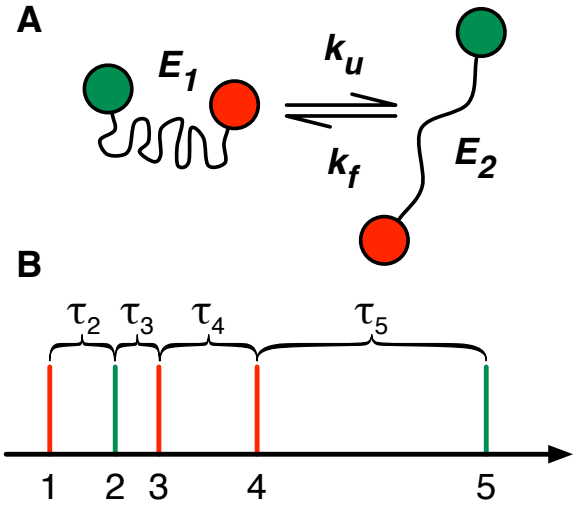

Figure 3.10: Illustration of the GSML method. A: The GSML method as used in this work assumes a simple two-state model of unfolded and native protein with distinct transfer efficiencies E_1 and E_2, which interconvert with the rates k_u and k_f. B: An example of a five-photon burst: red – green – red – red - green. The likelihood for this burst according to equation 3.6 is $L = \mathbf{1}^T(\mathbf{I}-\mathbf{E})e^{\mathbf{K}\tau_5}\mathbf{E}e^{\mathbf{K}\tau_4}\mathbf{E}e^{\mathbf{K}\tau_3}(\mathbf{I}-\mathbf{E})e^{\mathbf{K}\tau_2}\mathbf{E}\mathbf{p}_{eq}$

to four hours or more. We analyzed the data presented in the histograms in figure 3.9 using the GSML approach as is summarized in figure 3.11. Only bursts with $E > 0.4$ were taken into account, so a description of the donor-only in terms of the model for the GSML method was not necessary.

The increase of $k_{tot} = k_u + k_f$ in figure 3.11A is in accordance with our expectation, that the folding kinetics become faster with increasing temperature. We see values of up to 6 ms^{-1}, which is faster than the diffusion time through the confocal volume. This signal averaging, which is faster than the observation timescale, would explain the emergence of a single peak in the transfer efficiency histograms with an average transfer efficiency between the two initial ones.

Despite the shape of the histogram being dictated by fast interconversion dynamics, one expects the transfer efficiency of the native state to remain constant, while the transfer efficiency of the unfolded state will change in a temperature dependent manner as described in chapter 2. The fitted transfer efficiencies of the unfolded state do not change much over the temperature range investigated, much less than the collapse of the P6C/I84C variant, this is expected due to the presence of GdmCl.

It is expected for a system near the folding midpoint that the equilibrium constant $K = k_f/k_u$ is approximately one. Increasing the temperature should result in unfolding of the protein, in turn this will decrease the the equilibrium constant. This behavior is not conclusively

shown in the data. Instead the equilibrium constant varies in a non-continuos fashion.

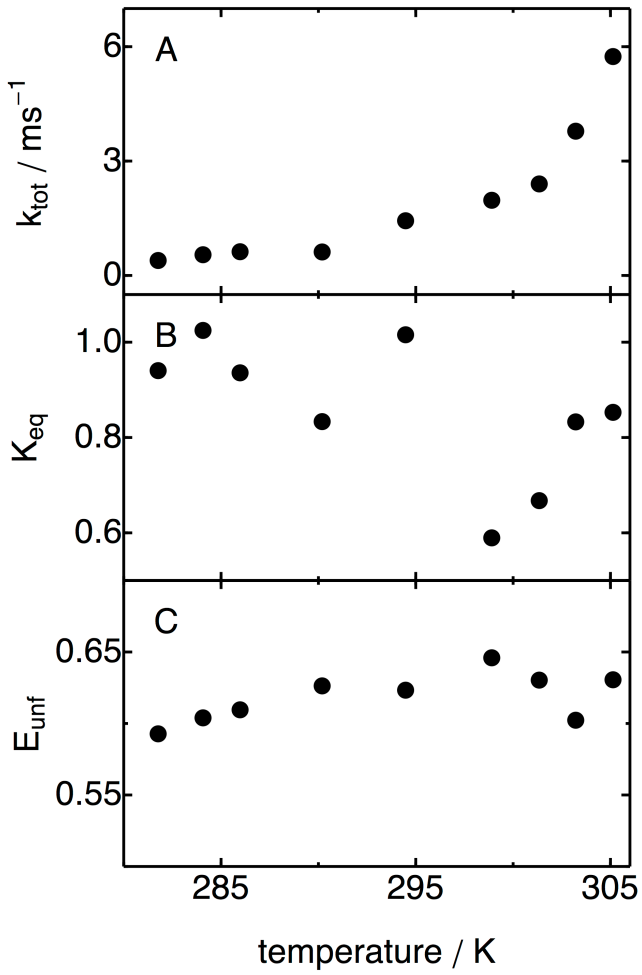

Figure 3.11: Fitted GSML parameters Fitted parameters using the GSML method on the λ repressor measurement of figure 3.9. **A:** sum of folding and unfolding rate constants k_{tot}; **B** Equilibrium constant; **C** transfer efficiency of the unfolded fitted form the data

In order to asses the quality of these fits, we showed that a set of simulated bursts generated by using the determined kinetic parameters can reproduce the histograms. We do this by applying the recoloring method, as outlined in Gopich and Szabo 2009. To do so we retain the arrival times of all photons in all bursts and disregard their color. Then the bursts are recolored according to the fitted parameters. These simulated bursts are used to construct FRET efficiency histograms, which then can be compared to the observed ones, as seen in figure 3.12. Experimental and simulated histograms are in good agreement in the case of the λ-repressor variant T8C/K70C.

Although the the simulated histograms describe the measured ones well, some doubt remained due to the behavior of the equilibrium constant. For testing the reliability of the GSML

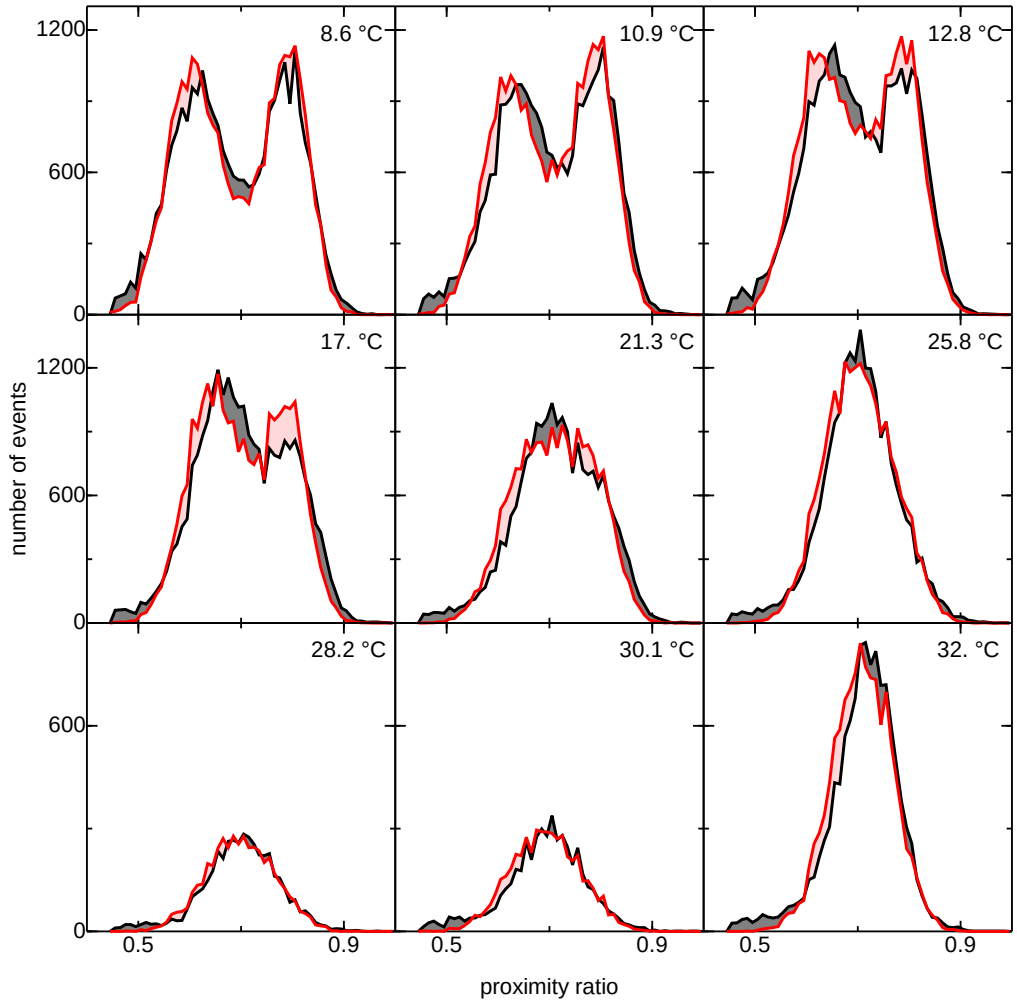

Figure 3.12: Temperature dependent data vs. simulated data. Proximity ratio (uncorrected FRET efficiency) histograms for measurements at different temperatures as shown in figure 3.9 in black vs. simulated data in red.

method, we applied it to Csp*Tm*, a protein model system where the folding rates are know. We explicitly used Csp*Tm*, as an example for a protein with a lack of sub-millisecond folding dynamics. We analyzed Csp*Tm* data near its unfolding midpoint at 1.2 M GdmCl at room temperature as shown in figure 3.13. We obtain a k_{tot} of 311 s^{-1}, which is much faster than the known rates from stopped flow experiments (Schuler *et al.*, 2002) in the 1 s^{-1} regime. If we set $k_{tot} = 10^{-6}\mu s^{-1}$ in conjunction with the fitted transfer efficiencies, we obtain recolored histograms that do not fit the experimental data. So far it is not clear what causes this overestimation of the rates. We speculate that the diffusion timescale still influences the photon statistics because some violation of the fundamental assumptions, photon colors being independent of the position of the protein in the confocal volume or the total count rate be-

ing independent of the conformation, do not hold completely, maybe due to some chromatic aberration of the used optical elements.

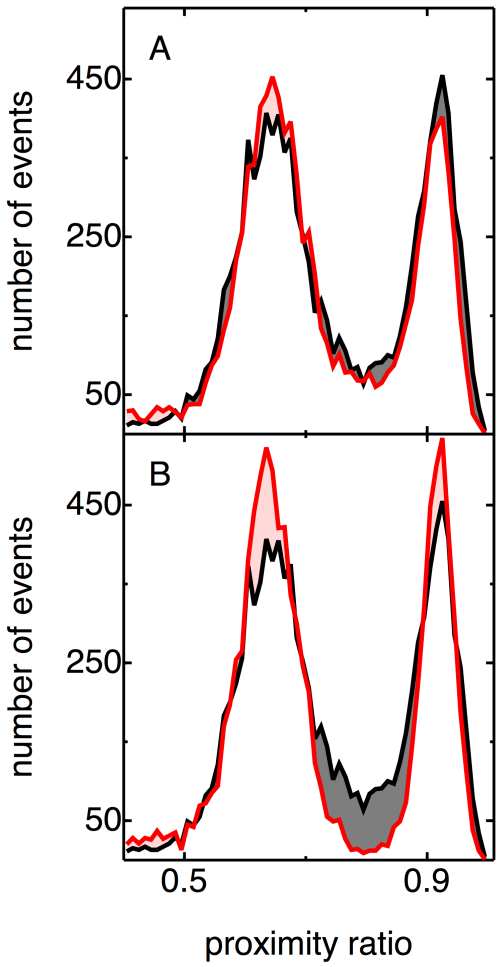

Figure 3.13: GSML method applied on Csp*Tm* A: Analysis on Csp*Tm* data, similar to figure 3.12. The experimental data is shown in black, the simulated recolored data is shown in red. **B:** The same experimental data as above, but recoloring performed not with the obtained rates, but a more realistic fixed value of $k_{tot} = 1s^{-1}$.

However, a second interpretation of the results is possible. Just as with Csp*Tm* (Nettels *et al.*, 2009) and λ repressor P6C/I84C, in chapter 2 of this thesis, a scenario is possible in which with increasing temperature the folded population decreases and the unfolded ensemble collapses to an average end-to-end distance, which coincidentally corresponds to the mean of the initial transfer efficiencies, solely explaining the shape of the histograms with temperature induced collapse. Some preliminary experiments on T8C/R82C support this, as they show a similar extent of change in transfer efficiency in the unfolded state over temperature,

as seen in figure 3.14. The GSML method might not be robust enough to extract folding rate constants between populations of changing transfer efficiency or only model systems with vastly different folding times. Recurrence analysis of single particles (Hoffmann *et al.*, 2011), as an alternative way to analyze burst trajectories, has shown its ability to work over several different time scales and will be later used to confirm the lack of fast dynamics on the P6C/I84C variant. Since the three variants only differ in the position of dye attachment, it seems likely that, also with T8C/K70C, we are looking at temperature induced collapse of the unfolded state instead of fast interconversion dynamics.

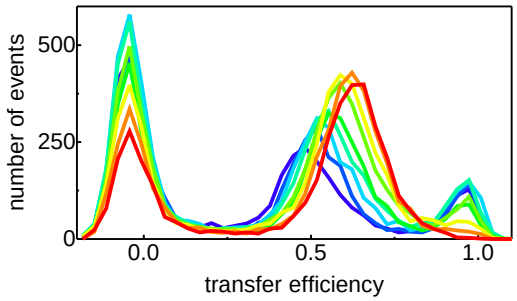

Figure 3.14: Temperature dependent histograms of T8C/R82C Histograms of T8C/R82C in 0.73 M GdmCl (similar to the conditions in figure 3.9) from 8.6 (blue) to 32°C (red), showing both the temperature collapse of the unfolded state and the continuous unfolding.

3.4.3 Recurrence analysis of single particles

A standard way to analyze data from single-molecule FRET experiments on freely diffusing molecules is to identify the fluorescence bursts and produce transfer efficiency histograms. Usually bursts are treated as independent events that all stem from different molecules. However, even visual inspection of typical signal traces already suggests that bursts are not completely independent of each other. In trajectories, bursts often appear in clusters, as can be seen in figure 3.15. Due to the very low concentration in the tens of picomolar range, the probability of two subsequent bursts being from the same molecule reentering the focus, instead of originating from two different molecules, is high when the time interval between the two bursts is short. These reentering molecules also lead to a contribution to the correlation function in FCS experiments at times longer than the mean diffusion time through the confocal volume. The recurrence effect can be used to extract kinetic information, even on timescales longer than the burst duration, as has been shown by Hoffmann *et al.* 2011. In their study, the folding dynamics of three different proteins have been analyzed: Csp*Tm*, correctly showing no dynamics on the recurrence timescale; spectrin R15, showing relaxation kinetics

around 31 ms^{-1} and B domain of protein A showing a folding rate of 0.93 ms^{-1}. This makes RASP suitable for the search for fast folding dynamics on the timescale of diffusion. Another advantage of RASP is that it is largely model-free.

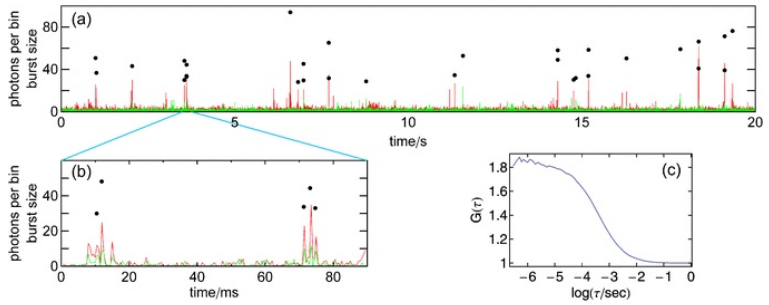

Figure 3.15: RASP example trajectory The recurrence effect in single-molecule FRET experiments of freely diffusing molecules is clearly visible as clustering of photon bursts in binned time traces (a,b) (donor signal in green, acceptor signal in red, 1 ms binning). Points indicate the positions and the numbers of photon counts of identified bursts. The clustering is also the origin of fluorescence intensity correlation on time scales greater than the mean diffusion time (B1 ms) through the confocal volume (c). Data were measured on freely diffusing FRET-labeled CspTm in 1.1 M GdmCl. Taken from Hoffmann *et al.* 2011

The derivation of RASP with the underlying mathematics and necessary corrections is presented in the original paper (Hoffmann *et al.*, 2011), but here I will describe the two basic components. The first component is the recurrence histogram. Recurrence histograms are histograms that are constructed from a subset of the bursts of the measured data. The bursts are selected by two criteria: They must be detected after a first burst that has a transfer efficiency within a certain range - the initial E-range, e.g. 0.5-0.6. Also, they must be detected in a certain time span, were τ is the mean time of the interval, after the first burst. All bursts in a measurement that satisfy these criteria are used to construct the recurrence histogram. Three examples for such histograms are shown in figure 3.16, with the initial E-ranges shown in red boxes and τ range of 0-4.1 ms.

Now consider the recurrence effect: Simplified, this means that at short recurrence times the recurrence histograms are approximately histograms of proteins reentering the focus. For example if one selects a transfer efficiency corresponding to the unfolded state, the recurrence histogram will mostly show an unfolded peak, with some residual contribution of new molecules. However, in case of fast folding dynamics on the timescale of the recurrence time, a peak with native state transfer efficiency will emerge, corresponding to proteins that have folded. Essentially one can use the population of the folded state in the recurrence histograms in conjunction with the recurrence time to extract the folding rate constant. In order to do so, one generates a series of recurrence histograms with different τ-intervals. Some examples

of such histograms are shown in figure 3.18[4]. However, there is a residual probability that the bursts do not originate from the same but a new molecule entering the focus. In order to quantify the interconversion rate constants, one needs to account for this effect, when e.g. fitting the change in the fraction of unfolded protein in the series of recurrence histograms. Consider a two-state system, A and B. We define $p_A(\tau, \Delta E_1)$ as the probability for a set of burst pairs , where the initial burst had a transfer efficiency within ΔE_1 and the second burst was detected at τ, that the second bursts is emitted by a molecule in state A. This molecule in state A can either be a recurring ($i = j$) or a new molecule ($i \neq j$) entering the focus. Therefore $p_A(\tau, \Delta E_1)$ can be partitioned accordingly:

$$p_A(\tau, \Delta E_1) = p_{\text{same}}(\tau) p_A^{i=j}(\tau, \Delta E_1) + [1 - p_{\text{same}}(\tau)]\, p_A^{i \neq j} \tag{3.10}$$

$p_A(\tau, \Delta E_1)$ can be experimentally determined from fitting the corresponding recurrence histogram and determining the ratio of the peak area corresponding to subpopulation A over the total area under the peaks corresponding to A and B. $p_A^{i=j}(\tau, \Delta E_1)$ and $p_A^{i \neq j}$ are the probabilities that a recurring or a new molecule are in state A, respectively . The latter is given by the equilibrium probability, which can be determined from the full transfer efficiency histogram. $p_{\text{same}}(\tau)$ is the probability that both bursts originate from the same molecule ($i = j$). The time dependence of $p_A^{i=j}(\tau, \Delta E_1)$ can now be used to extract conformational dynamics:

$$p_A^{i=j}(\tau, \Delta E_1) = \rho_A^{eq} + [\rho_A(0, \Delta E_1) - \rho_A^{eq}]\, e^{-k_{tot}\tau} \tag{3.11}$$

ρ_A^{eq} is the equilibrium probability that a protein in state A. $\rho_A(0, \Delta E_1)$ is the initial probability that a protein that emitted a burst at time 0 with a transfer efficiency in the range ΔE_1 is in state A at time $\tau = 0$. $p_A^{i=j}(\tau, \Delta E_1)$ probability evolves over time with the rate constant $k_{tot} = k_f + k_u$.

The last component necessary to determine this rate constant is the probability that bursts are emitted by the initial molecule. Bursts from different proteins should be uncorrelated in time, while bursts originating from the same should correlate. We define a burst time correlation in analogy to photon time correlation in FCS:

$$g(\tau) = \frac{p(t, t+\tau)}{p(t)p(t+\tau)} \tag{3.12}$$

where $p(t, t+\tau)$ is the joint probability of detecting a burst at time τ after first detecting a burst at time t. $p(t)$ and $p(t+\tau)$ are the independent probabilities of detecting bursts at time points t and $t+\tau$, respectively, where t is the temporal midpoint of the burst. Hoffmann *et al.* showed that the 'same molecule probability' is given by:

$$p_{\text{same}}(\tau) = 1 - \frac{1}{g(\tau)} \tag{3.13}$$

[4]As will be discussed later, these histograms are not influenced by fast interconversion dynamics.

An example for τ-dependent same molecule probability for a measurement on λ repressor is shown in figure 3.17. The τ dependency of p_{same} was fit empirically according to Hoffmann *et al.* 2011. Then τ with $p_{\text{same}} > 0.95$ was determined (4.1 ms) and used to construct the series of recurrence histograms in figure 3.16.

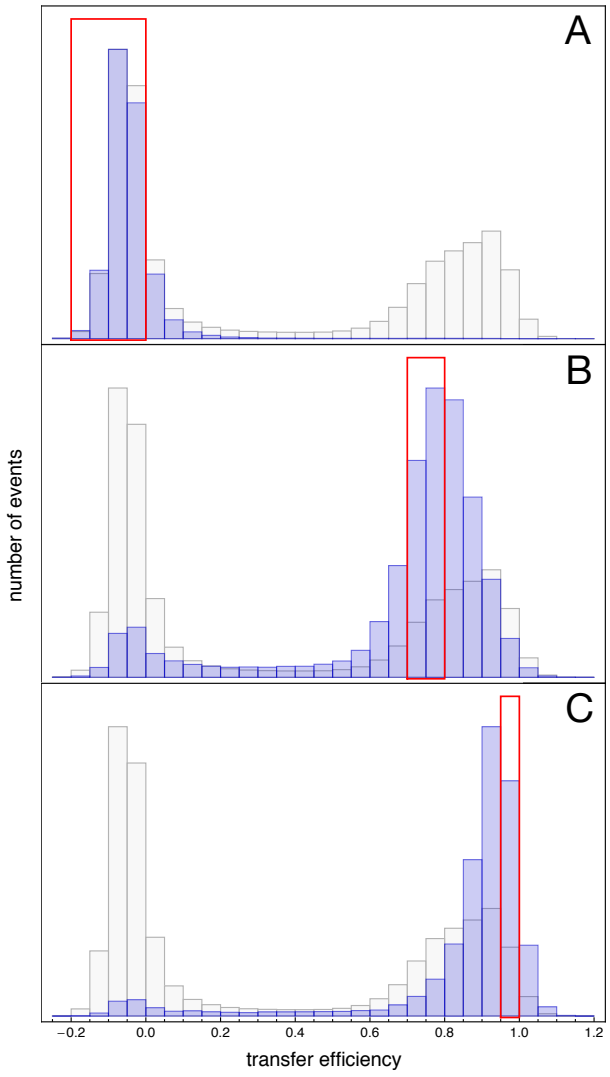

Figure 3.16: Recurrence histograms Recurrence histograms of λ repressor P6C I84C at 298 K in native solvent conditions. Recurrence histograms of all bursts recurring after the initial burst within 4.1 ms ($p_{same} \geq 0.95$) are used. Shown are normalized recurrence histograms in blue in the background of the full histogram. The initial E-range is marked as a red box. Therefore the panels correspond, top to bottom, to predominantly the donor-only fraction, the unfolded state and the native state, respectively.

The outlined procedure now can be used to extract folding kinetics from single molecule equilibrium experiments. The final result of the RASP analysis is shown in figure 3.19. This shows the fraction of unfolded molecules dependent on the recurrence time interval for different initial transfer efficiency ranges (unfolded and native). The colored lines are fits to the

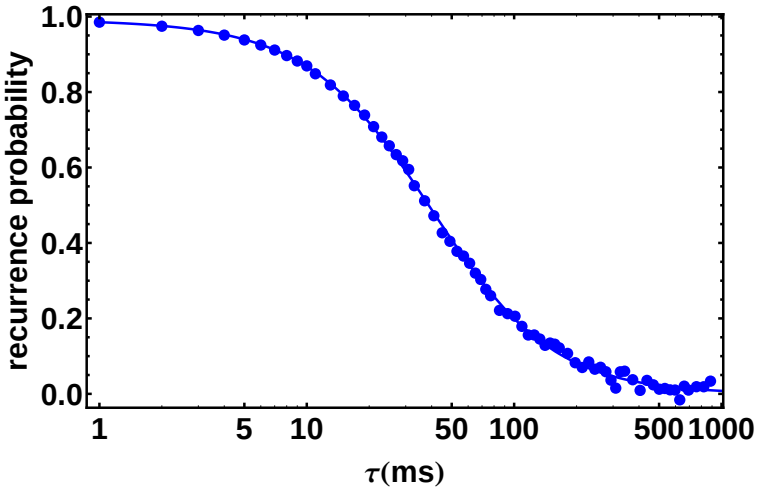

Figure 3.17: Same molecule probability Same molecule probability $p_{\text{same}}(\tau)$ for a measurement of λ repressor P6C/I84C in native buffer at 298 K. The line is an empirical fit according to $p_{\text{same}}(\tau) = 1 - \frac{1}{1+n^{-1}(1+\tau/\tau_D)^{-3/2}}$ (Hoffmann *et al.*, 2011).

data according to equation 3.10. The dashed lines are the change in populations one would expect solely from the appearance or new molecules ($\lambda = 0$). Both curves are indistinguishable. Therefore, similar to Csp*Tm*, λ repressor P6C/I84C shows no folding dynamics on the recurrence timescale at 298 K. The goal of this particular experiment was to test the lack of fast conformational dynamics in context of the studies in the first part of the thesis. Since all three λ repressor variants are essentially the same and only differ in the position of the dye attachment, one would have expected some fast dynamics. This however is not reflected in the data.

3.5 Conclusion and outlook

3.5.1 Related studies

The initial aim of this project to measure the timescale and possibly resolve the process of barrier transition in the protein folding reaction remained unsuccessful. As a first step would be to meaningful relate the rate of folding with the timescale of diffusion through the confocal volume. These need to be in the same range. For this purpose one needs to determine the unfolding and folding rate coefficients of the proteins. Ideally one wants to extract these rate coefficients from the same single-molecule equilibrium experiments that are needed to obtain the distributions of pathways which lead to folding. The GSML approach holds the promise to extract the rate coefficients and possibly give some indication for the temporal position of

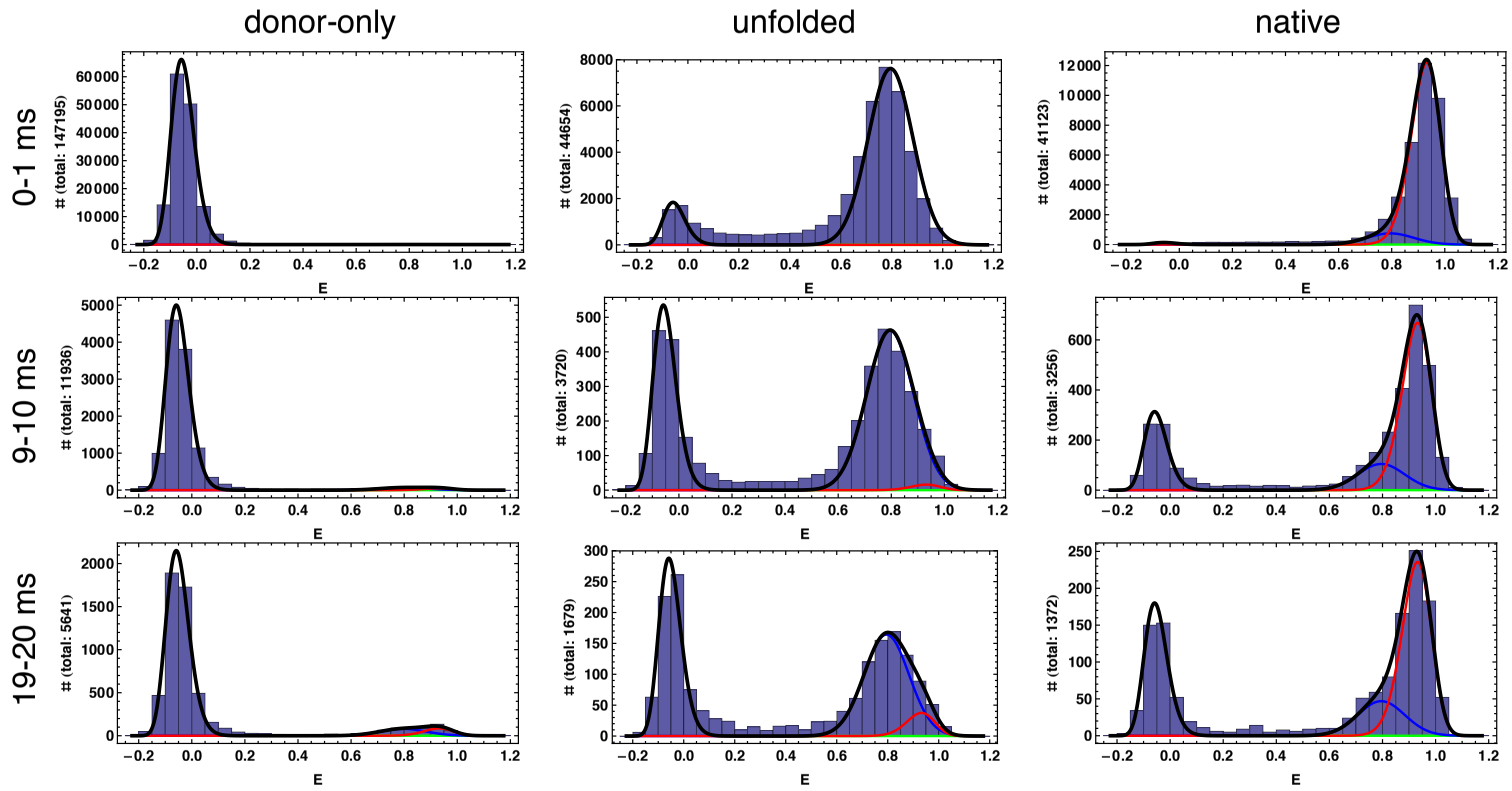

Figure 3.18: Recurrence time window dependent histograms These are recurrence histograms generated from the same dataset as in figure 3.16. Instead of producing histograms from all recurring burst with a certain p_{same} threshold, we generate histograms from bursts in a certain recurrence time window (0-1, 9-10, 19-20 ms in this example) after a burst with a certain initial E-range corresponding to the histogram peaks (donor-only, unfolded,native).

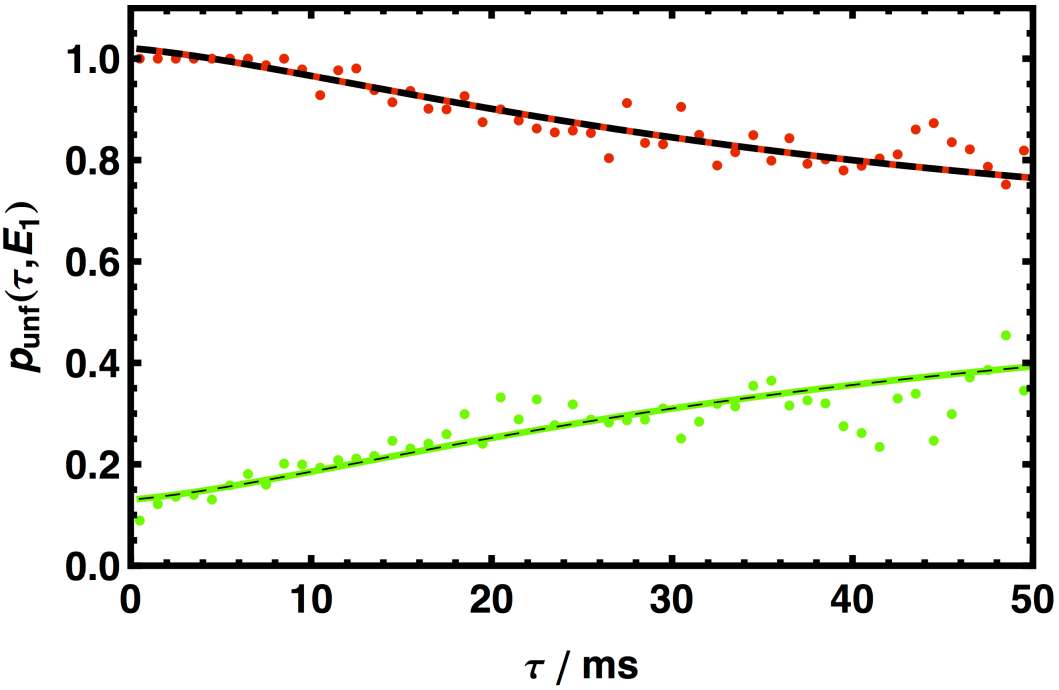

Figure 3.19: Recurrence kinetics of λ repressor P6C I84C at 298 K in native solvent conditions Shown is the fraction of unfolded molecules at certain recurrence times starting out from the unfolded (red) or the folded (green) population. The solid lines are a global fit to equation 3.10 describing folding and unfolding of recurring molecules. The dashed lines indicate the change on fraction folded as one would expect solely from the occurrence of new molecules. Since both lines superimpose, this illustrates that no interconversion dynamics influence the recurrence histograms.

the folding reaction. Over the course of this PhD work the group of William Eaton (Chung *et al.*, 2009, 2011, 2012) investigated the folding of several proteins in the the single-molecule FRET context including successfully applying the GSML method. They analyzed trajectories for the fast folding, three-helix two-state artificial protein α_3D, the all-β WW domain peptide, and the α/β protein GB1. In their first study they observed α_3D, both in free diffusion and immobilized on polyethylenglycol coated glass. Even a simple analysis based on FRET efficiency histograms showed some striking results: On FRET efficiency histograms of protein at the chemical unfolding midpoint, exhibiting two easily distinguishable populations, the authors could vary the bin time of photons trajectory, used for generating the histograms. They observed a gradual change from two populations at high and low transfer efficiency to a single population at an intermediate transfer efficiency as they increased the bin time. This has been found with data on immobilized proteins, the early stages of this development however were also exhibited on free diffusing proteins, where the bin time is limited by the diffusion time. This is already a strong indicator for fast folding dynamics. The authors were able to determine interconversion rates by fitting the histograms with three gaussian distributions and

relating their populations to the bin time. Furthermore, they applied the GSML method to the data, and find good agreement of the rate coefficients between experiments on free diffusing and immobilized proteins and the two analysis methods based on histograms or the GSML method. These analysis methods yielded observable rate constants in the low millisecond range $k_{obs} = 1\text{-}2\, ms^{-1}$. Chung *et al.* even analyzed the trajectories in search for transition time points using the Viterbi algorithm (Viterbi, 1967), a dynamic programming algorithm for finding the most likely sequence of hidden states, that results in a sequence of observed events. However, they assume that a photon either belongs to the unfolded or the folded state make the transition instantaneous. In more recent work, Chung *et al.* 2012 adapted the GSML method to also determine the transition path time. They do so in expanding the two-state kinetic model by a state along the transition path, with a transfer efficiency midway between the unfolded and the native state. The lifetime of this state corresponds to the average transition path time. They determine a transition path time for WW in 3 M GdmCl, 50% glycerol of 16 µs, and extrapolate it to native-like conditions at 2 µs. For GB1, they determined an upper bound for the transition path time of approx. 10 µs. The determined transition path times were insensitive to the exact position an the number of intermediate transfer efficiency steps between the unfolded and the native state. It is remarkable that, although the proteins differ in folding times by several orders of magnitude, the average transition path time for GB1 is only five times longer than for the WW domain. The authors point out that for a reliable extraction of rate coefficients and transition path times a detailed understanding of the photon trajectories is paramount. This ultimately leads to the exclusion of trajectories from the analysis of conformational dynamics. Dye photophysics, e.g. a red shifted state of the Alexa 488 dye, significantly affect the trajectories, as outlined by the authors of the above studies in a separate publication (Chung *et al.*, 2010).

3.5.2 Findings

My present work on the folding dynamics and the transition path times of λ repressor remains inconclusive. However, the difficulties encountered, against the backdrop of the studies by Eaton *et al.* presented above, might present valuable starting points in improving the approach. Circular dichroism data on unlabeled protein as well as FRET efficiency histograms and ensemble unfolding transitions of labeled protein indicate that the pseudo-wild type used in the thesis, in all his variations of labeling positions, is capable of adapting a native conformation. On which timescale the folding occurs was merely informed on by two pieces of data: The stopped-flow experiments on variant T8C/R82C and the GSML analysis of T8C/K70C, both pointing to millisecond timescales. On the other hand, stopped-flow experiments on T8C/K70C and recurrence-analysis data on P6C/I84C were not able to show any relaxation

at all. The reasons for this are not clear so far. For distinguishing different states in FRET experiments one wants to maximize their difference in average transfer efficiency - obviously both state have to be observable (bright). This certainly has been an issue in this work, e.g. making recurrence histogram hard to fit. Especially in comparison with the work by Eaton *et al.* the difference between unfolded and native state in our GSML-work on T8C/K70C was much smaller, making it more difficult to assign burst of immediate transfer efficiency to the emergence of fast interconversion dynamics. A small amplitude also limits the application of ensemble methods like stopped-flow rapid or microfluidic mixing.

In conjunction with the GSML method, we encountered another important pitfall: the lack of a negative control for the GSML method in the context of free diffusing molecules, as Csp*Tm* also showed millisecond folding times. Since Csp*Tm* folding time is in the seconds range, and Eaton *et al.* could successfully extract slow rate coefficients from GB1 in an immobilized setup, we hypothesize that the diffusion process through the confocal volume "leaks" into the interconversion process. Also the equilibrium constant did not change in a continuous fashion, out of line with the expected population shift by unfolding as temperature is increased.

3.5.3 Outlook

From this situation, one can identify several aims to further develop the study of protein folding rate coefficients and transition paths. One possible approach would be to increase the transfer efficiency separation between the unfolded and the native state. Ensemble methods on labelled proteins, like stopped-flow rapid mixing or temperature-jump would gain in available amplitude from an increased separation. In the context of analyzing single molecule data in terms of histograms, a greater signal separation will make it easier to distinguish between peak broadening due to shotnoise or fast dynamics, especially in histograms comprised only of a few bursts, like recurrence histograms at long time scales. Also the immediately intuitive consequence of fast dynamics on bin-size dependent histograms in the emergence of populations of an average transfer efficiency so far only has been become apparent the works of Chung and Eaton (Chung *et al.*, 2009), where a separation of 0.3 in transfer efficiency was available.

A greater difference in transfer efficiency can be achieved by choosing appropriate sites for introducing the cysteine residues to facilitate dye labeling. In an experiment that probes the overall protein compactness, one will in most cases want to attach the dyes near the termini of the chain. In the native state, however the dyes should be close together to maximize the difference in transfer efficiency. This in turn might lead to the problem that the dyes quench each other, which apparently was the case in the T8C/R82C variant. The rather dark acceptor

in this variant might very well be caused by quenching of the donor. Unpublished results by Hänni *et al.* hint to the possibility of this. Since direct contact is required for quenching, one should place the dyes close to one another in the native state but somewhat shielded from each other by the protein. This was attempted in the T8C/K70C variant, which in turn diminished the difference between unfolded and native state transfer efficiency again.

An alternative approach, would be to utilize different, more stable model proteins. In this case, reaching the unfolding midpoint would demand a higher concentration of denaturant, leading to a more expanded unfolded state resulting in a bigger difference in transfer efficiency between the unfolded and the native state. Histograms and recurrence data on the B domain of protein A presented in Hoffmann *et al.* 2011, already suggest that it might be a more suitable candidate.

Free diffusion experiments only have very short observation times available, essentially limiting the time scale window in which folding rates can be determined. The protein has to be engineered/selected in way to match its inverse diffusion time through the confocal volume. This ensures that enough proteins undergo a (un)folding during the burst time, so that they influence histogram shapes. Since the diffusion time is in the millisecond range, it is quite hard to achieve this time scale at the "slowest" region of the chevron plot. Ideally it even should fulfill the demand to follow a downhill folding scenario, which is opposed to the demand that the solvent conditions should be near to the midpoint.

3.5.4 Further steps

So far folding transitions only could easily be observed in the context of single molecules which were immobilized on a surface and observed for extended periods of time (several milliseconds and more). Immobilizing molecules has the obvious advantage of allowing longer observation times, and thereby to longer continuous trajectories. This potentially allows easier analysis, since identification of states benefits from longer signal traces. Therefore it might be beneficial to explore options for immobilization. However such an experimental setup posed the problem of the influence of the surface on the protein (and dyes). These problems vary in degree and type depending on the coupling chemistry and can effect protein structure and kinetics and/or fluorophore behavior. How this affects folding trajectories is often neglected, even though detailed understanding is essential for the extraction of quantitative information (Chung *et al.*, 2010). Some ways to mitigate problems of immobilization would be the embedding of the protein into tethered lipid vesicles (Rhoades *et al.*, 2004) or into a gel matrix (Santoso and Kapanidis, 2009), however this adds additional experimental complications.

A problem which both experimental setups have to deal with is the relatively low photon

rate, compared the expected timescale of the transition path time. The photon rate effectively limits the time resolution achievable in FRET experiments. Even the works of Chung and Eaton were not capable of further resolving the folding pathway. However, the development of new dyes and complimentary "photoprotective" cocktails is a topic of ongoing research. These aim to reduce blinking, bleaching and residence in non-fluorescent "dark states" (triplet state). Many groups report improvements of chromophore properties by using redox systems (Vogelsang *et al.*, 2008; Cordes *et al.*, 2009; Liu *et al.*, 2012). Such developments might help to ultimately resolve the detail of protein folding transition paths.

Bibliography

Ackers G.K.; Johnson A.D.; Shea M.A. Quantitative model for gene regulation by lambda phage repressor. *Proceedings of the National Academy of Sciences of the United States of America*, **79**(4):1129–1133 (1982).

Arora P.; Oas T.G.; Myers J.K. Fast and faster: a designed variant of the B-domain of protein A folds in 3 microsec. *Protein science : a publication of the Protein Society*, **13**(4):847–853 (2004).

Beamer L.J.; Pabo C.O. Refined 1.8 A crystal structure of the lambda repressor-operator complex. *Journal of Molecular Biology*, **227**(1):177–196 (1992).

Brinkley M. A brief survey of methods for preparing protein conjugates with dyes, haptens, and cross-linking reagents. *Bioconjugate chemistry*, **3**(1):2–13 (1992).

Bryngelson J.D.; Onuchic J.N.; Socci N.; Wolynes P. Funnels, Pathways and the Energy Landscape of Protein Folding: A Synthesis. *Proteins: Structure, Function, and Bioinformatics*, **21**(3):167–195 (1995).

Burton R.E.; Huang G.S.; Daugherty M.A.; Calderone T.L.; Oas T.G. The energy landscape of a fast-folding protein mapped by Ala–>Gly substitutions. *Nature Structural Biology*, **4**(4):305–310 (1997).

Burton R.E.; Huang G.S.; Daugherty M.A.; Fullbright P.W.; Oas T.G. Microsecond protein folding through a compact transition state. *Journal of Molecular Biology*, **263**(2):311–322 (1996).

Chugha P.; Oas T.G. Backbone dynamics of the monomeric lambda repressor denatured state ensemble under nondenaturing conditions. *Biochemistry*, **46**(5):1141–1151 (2007).

Chugha P.; Sage H.J.; Oas T.G. Methionine oxidation of monomeric lambda repressor: the denatured state ensemble under nondenaturing conditions. *Protein science : a publication of the Protein Society*, **15**(3):533–542 (2006).

Chung H.S.; Gopich I.V.; McHale K.; Cellmer T.; Louis J.M.; Eaton W.A. Extracting Rate Coefficients from Single-Molecule Photon Trajectories and FRET Efficiency Histograms for a Fast-Folding Protein. *The Journal of Physical Chemistry A*, **115**(16):3642–3656 (2011).

Chung H.S.; Louis J.M.; Eaton W.A. Experimental determination of upper bound for transition path times in protein folding from single-molecule photon-by-photon trajectories. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(29):11837–11844 (2009).

Chung H.S.; Louis J.M.; Eaton W.A. Distinguishing between Protein Dynamics and Dye Photophysics in Single-Molecule FRET Experiments. *Biophysical Journal*, **98**(4):696–706 (2010).

Chung H.S.; McHale K.; Louis J.M.; Eaton W.A. Single-Molecule Fluorescence Experiments Determine Protein Folding Transition Path Times. *Science*, **335**(6071):981–984 (2012).

Cobos E.S.; Radford S.E. Sulfate-induced effects in the on-pathway intermediate of the bacterial immunity protein Im7. *Biochemistry*, **45**(7):2274–2282 (2006).

Cordes T.; Vogelsang J.; Tinnefeld P. On the mechanism of Trolox as antiblinking and antibleaching reagent. *Journal of the American Chemical Society*, **131**(14):5018–5019 (2009).

Doering D. . Ph.D. thesis, Massachusetts Institute of Technology (1992).

Dumont C.; Matsumura Y.; Kim S.J.; Li J.; Kondrashkina E.; Kihara H.; Gruebele M. Solvent-tuning the collapse and helix formation time scales of lambda(6-85)*. *Protein science : a publication of the Protein Society*, **15**(11):2596–2604 (2006).

Eaton W.A. Searching for "downhill scenarios" in protein folding. *Proceedings of the National Academy of Sciences of the United States of America*, **96**(11):5897–5899 (1999).

Gambin Y.; VanDelinder V.; Ferreon A.C.M.; Lemke E.A.; Groisman A.; Deniz A.A. Visualizing a one-way protein encounter complex by ultrafast single-molecule mixing. *Nature Methods*, **8**(3):239–241 (2011).

Garcia-Mira M.M.; Sadqi M.; Fischer N.; Sanchez-Ruiz J.M.; Muñoz V. Experimental identification of downhill protein folding. *Science*, **298**(5601):2191–2195 (2002).

Ghaemmaghami S.; Word J.M.; Burton R.E.; Richardson J.S.; Oas T.G. Folding kinetics of a fluorescent variant of monomeric lambda repressor. *Biochemistry*, **37**(25):9179–9185 (1998).

Godoy-Ruiz R.; Henry E.R.; Kubelka J.; Hofrichter J.; Muñoz V.; Sanchez-Ruiz J.M.; Eaton W.A. Estimating free-energy barrier heights for an ultrafast folding protein from calorimetric and kinetic data. *The Journal of Physical Chemistry B*, **112**(19):5938–5949 (2008).

Gopich I.V.; Szabo A. Single-molecule FRET with diffusion and conformational dynamics. *The Journal of Physical Chemistry B*, **111**(44):12925–12932 (2007).

Gopich I.V.; Szabo A. Decoding the Pattern of Photon Colors in Single-Molecule FRET. *The Journal of Physical Chemistry B*, **113**(31):10965–10973 (2009).

Gruebele M.; Sabelko J.; Ballew R.; Ervin J. Laser temperature jump induced protein refolding. *Accounts of Chemical Research*, **31**(11):699–707 (1998).

Ha T.; Tinnefeld P. Photophysics of fluorescent probes for single-molecule biophysics and super-resolution imaging. *Annual Review of Physical Chemistry*, **63**:595–617 (2012).

Hamadani K.M.; Weiss S. Non-equilibrium Single Molecule Protein Folding in a Co-axial Mixer. *Biophysical Journal* (2008).

Hänggi P.; Talkner p.; Borkovec M. Reaction-rate theory: fiftyl years after Kramers. *Reviews in Modern Physics*, **62**:251–341 (1990).

Hillger F.; Hänni D.; Nettels D.; Geister S.; Grandin M.; Textor M.; Schuler B. Probing protein-chaperone interactions with single-molecule fluorescence spectroscopy. *Angewandte Chemie (International ed in English)*, **47**(33):6184–6188 (2008).

Hoffmann A.; Kane A.; Nettels D.; Hertzog D.E.; Baumgartel P.; Lengefeld J.; Reichardt G.; Horsley D.A.; Seckler R.; Bakajin O.; Schuler B. Mapping protein collapse with single-molecule fluorescence and kinetic synchrotron radiation circular dichroism spectroscopy. *Proceedings of the National Academy of Sciences of the United States of America*, **104**(1):105–110 (2007).

Hoffmann A.; Nettels D.; Clark J.; Borgia A.; Radford S.E.; Clarke J.; Schuler B. Quantifying heterogeneity and conformational dynamics from single molecule FRET of diffusing molecules: recurrence analysis of single particles (RASP). *Phys. Chem. Chem. Phys.*, **13**(5):1857–1871 (2011).

Hofmann H.; Hillger F.; Pfeil S.H.; Hoffmann A.; Streich D.; Haenni D.; Nettels D.; Lipman E.A.; Schuler B. Single-molecule spectroscopy of protein folding in a chaperonin cage. *Proceedings of the National Academy of Sciences of the United States of America* (2010).

Hohng S.; Joo C.; Ha T. Single-molecule three-color FRET. *Biophysj*, **87**(2):1328–1337 (2004).

Huang G.S.; Oas T.G. Structure and stability of monomeric lambda repressor: NMR evidence for two-state folding. *Biochemistry*, **34**(12):3884–3892 (1995a).

Huang G.S.; Oas T.G. Submillisecond folding of monomeric lambda repressor. *Proceedings of the National Academy of Sciences of the United States of America*, **92**(15):6878–6882 (1995b).

Hummer G. From transition paths to transition states and rate coefficients. *The Journal of Chemical Physics*, **120**:516 (2004).

Kramers H. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, **7**:284–304 (1940).

Kubelka J.; Hofrichter J.; Eaton W.A. The protein folding 'speed limit'. *Current Opinion in Structural Biology*, **14**(1):76–88 (2004).

Lee J.; Lee S.; Ragunathan K.; Joo C.; Ha T.; Hohng S. Single-molecule four-color FRET. *Angewandte Chemie (International ed in English)*, **49**(51):9922–9925 (2010).

Lei H.; Deng X.; Wang Z.; Duan Y. The fast-folding HP35 double mutant has a substantially reduced primary folding free energy barrier. *The Journal of Chemical Physics*, **129**(15):155104 (2008).

Li P.; Oliva F.Y.; Naganathan A.N.; Muñoz V. Dynamics of one-state downhill protein folding. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(1):103–108 (2009).

Lim W.A.; Hodel A.; Sauer R.T.; Richards F.M. The crystal structure of a mutant protein with altered but improved hydrophobic core packing. *Proceedings of the National Academy of Sciences of the United States of America*, **91**(1):423–427 (1994).

Lindorff-Larsen K.; Piana S.; Dror R.O.; Shaw D.E. How fast-folding proteins fold. *Science*, **334**(6055):517–520 (2011).

Lipman E.A.; Schuler B.; Bakajin O.; Eaton W.A. Single-molecule measurement of protein folding kinetics. *Science*, **301**(5637):1233–1235 (2003).

Liu F.; Du D.; Fuller A.A.; Davoren J.E.; Wipf P.; Kelly J.W.; Gruebele M. An experimental survey of the transition between two-state and downhill protein folding scenarios. *Proceedings of the National Academy of Sciences of the United States of America*, **105**(7):2369–2374 (2008).

Liu F.; Gruebele M. Tuning lambda(6-85) Towards Downhill Folding at its Melting Temperature. *Journal of Molecular Biology*, **370**(3):574–584 (2007).

Liu J.; Campos L.A.; Cerminara M.; Wang X.; Ramanathan R.; English D.S.; Muñoz V. Exploring one-state downhill protein folding in single molecules. *Proceedings of the National Academy of Sciences of the United States of America*, **109**(1):179–184 (2012).

Ma H.; Gruebele M. Kinetics are probe-dependent during downhill folding of an engineered lambda6-85 protein. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(7):2283–2287 (2005).

Meisner W.K.; Sosnick T.R. Barrier-limited, microsecond folding of a stable protein measured with hydrogen exchange: Implications for downhill folding. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(44):15639–15644 (2004).

Meyer B.J.; Kleid D.G.; Ptashne M. Lambda repressor turns off transcription of its own gene. *Proceedings of the National Academy of Sciences of the United States of America*, **72**(12):4785–4789 (1975).

Myers J.K.; Oas T.G. Contribution of a buried hydrogen bond to lambda repressor folding kinetics. *Biochemistry*, **38**(21):6761–6768 (1999).

Myers J.K.; Oas T.G. Preorganized secondary structure as an important determinant of fast protein folding. *Nature Structural Biology*, **8**(6):552–558 (2001).

Nettels D.; Gopich I.V.; Hoffmann A.; Schuler B. Ultrafast dynamics of protein collapse from single-molecule photon statistics. *Proceedings of the National Academy of Sciences of the United States of America*, **104**(8):2655–2660 (2007).

Nettels D.; Hoffmann A.; Schuler B. Unfolded Protein and Peptide Dynamics Investigated with Single-Molecule FRET and Correlation Spectroscopy from Picoseconds to Seconds. *The Journal of Physical Chemistry B* (2008).

Nettels D.; Müller-Späth S.; Küster F.; Hofmann H.; Haenni D.; Rüegger S.; Reymond L.; Hoffmann A.; Kubelka J.; Heinz B.; Gast K.; Best R.B.; Schuler B. Single-molecule spectroscopy of the temperature-induced collapse of unfolded proteins. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(49):20740–20745 (2009).

Noé F.; Schütte C.; Vanden-Eijnden E.; Reich L.; Weikl T.R. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(45):19011–19016 (2009).

Rhoades E.; Cohen M.; Schuler B.; Haran G. Two-state folding observed in individual protein molecules. *Journal of the American Chemical Society*, **126**(45):14686–14687 (2004).

Rhoades E.; Gussakovsky E.; Haran G. Watching proteins fold one molecule at a time. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(6):3197–3202 (2003).

Sabelko J.; Ervin J.; Gruebele M. Observation of strange kinetics in protein folding. *Proceedings of the National Academy of Sciences of the United States of America*, **96**(11):6031–6036 (1999).

Santoro M.M.; Bolen D.W. Unfolding free energy changes determined by the linear extrapolation method. 1. Unfolding of phenylmethanesulfonyl alpha-chymotrypsin using different denaturants. *Biochemistry*, **27**(21):8063–8068 (1988).

Santoso Y.; Kapanidis A.N. Probing biomolecular structures and dynamics of single molecules using in-gel alternating-laser excitation. *Analytical chemistry*, **81**(23):9561–9570 (2009).

Schuler B. Single-molecule fluorescence spectroscopy of protein folding. *ChemPhysChem*, **6**(7):1206–1220 (2005).

Schuler B.; Lipman E.A.; Eaton W.A. Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature*, **419**(6908):743–747 (2002).

Shastry M.C.; Luck S.D.; Roder H. A continuous-flow capillary mixing method to monitor reactions on the microsecond time scale. *Biophysical Journal*, **74**(5):2714–2721 (1998).

Spiegelman W.G.; Reichardt L.F.; Yaniv M.; Heinemann S.F.; Kaiser A.D.; Eisen H. Bidirectional transcription and the regulation of Phage lambda repressor synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, **69**(11):3156–3160 (1972).

Stayrook S.; Jaru-Ampornpan P.; Ni J.; Hochschild A.; Lewis M. Crystal structure of the lambda repressor and a model for pairwise cooperative operator binding. *Nature*, **452**(7190):1022–1025 (2008).

Streich D. Enhancing our understanding of larger protein folding: Rhodanese folding observed by single moleucle FRET spectroscopy. pages 1–34 (2010).

Thirumalai D.; O'Brien E.P.; Morrison G.; Hyeon C. Theoretical perspectives on protein folding. *Annual Review of Biophysics*, **39**:159–183 (2010).

Viterbi A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, **13**(2):260–269 (1967).

Vogelsang J.; Kasper R.; Steinhauer C.; Person B.; Heilemann M.; Sauer M.; Tinnefeld P. A reducing and oxidizing system minimizes photobleaching and blinking of fluorescent dyes. *Angewandte Chemie-International Edition*, **47**(29):5465–5469 (2008).

Yang W.Y.; Gruebele M. Folding at the speed limit. *Nature*, **423**(6936):193–197 (2003).

Yang W.Y.; Gruebele M. Folding lambda-repressor at its speed limit. *Biophysical Journal*, **87**(1):596–608 (2004).

Zwanzig R. Two-state models of protein folding kinetics. *Proceedings of the National Academy of Sciences of the United States of America*, **94**(1):148–150 (1997).

Chapter 4

Energy Surfaces from Single-Distance Information

I performed the single-molecule measurements on λ repressor including some preliminary data analysis, therefore providing the experimental data used with the FESST method.

Free Energy Surfaces from Single-Distance Information

Philipp Schuetz,[†] René Wuttke, Benjamin Schuler,* and Amedeo Caflisch*

Department of Biochemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland

Received: June 11, 2010; Revised Manuscript Received: September 21, 2010

We propose a network-based method for determining basins and barriers of complex free energy surfaces (e.g., the protein folding landscape) from the time series of a single intramolecular distance. First, a network of transitions is constructed by clustering the points of the time series according to the short-time distribution of the signal. The transition network, which reflects the short-time kinetics, is then used for the iterative determination of individual basins by the minimum-cut-based free energy profile, a barrier-preserving one-dimensional projection of the free energy surface. The method is tested using the time series of a single $C_\beta-C_\beta$ distance extracted from equilibrium molecular dynamics (MD) simulations of a structured peptide (20 residue three-stranded antiparallel β-sheet). Although the information of only one distance is employed to describe a system with 645 degrees of freedom, both the native state and the unfolding barrier of about 10 kJ/mol are determined with remarkable accuracy. Moreover, non-native conformers are identified by comparing long-time distributions of the same distance. To examine the applicability to single-molecule Förster resonance energy transfer (FRET) experiments, a time series of donor and acceptor photons is generated using the MD trajectory. The native state of the β-sheet peptide is determined accurately from the emulated FRET signal. Applied to real single-molecule FRET measurements on a monomeric variant of λ-repressor, the network-based method correctly identifies the folded and unfolded populations, which are clearly separated in the minimum-cut-based free energy profile.

I. Introduction

The thermodynamics and kinetics of a variety of complex systems, ranging from spin glasses to proteins, have been investigated by energy landscape theory in the 40 years since the publication of the seminal idea.[1] Peptides and proteins have a multidimensional and very complex potential energy surface with a large number of conformations of similar energy.[2,3] Yet, fast folding is possible because of the natural selection of sequences that make the native (i.e., functional) structure a pronounced energy minimum.[4] Entropic contributions are relevant at physiological temperatures, and therefore the *free* energy surface governs the thermodynamics and kinetics of polypeptide chains. In the past five years, new methods based on complex networks have been proposed to analyze free energy surfaces of folding,[5−10] which govern the process by which structured peptides or proteins assume their well-defined three-dimensional structure.

In view of the large number of microscopic folding pathways and the conformational heterogeneity in the denatured state, single molecule methods are a promising new approach to experimentally determine free energy surfaces.[11] One of the most versatile approaches, single molecule Förster resonance energy transfer (FRET), allows intramolecular distances and distance dynamics of individual protein molecules to be monitored.[12−17] Since distance distributions in different free energy states often overlap, the separability of the different basins is not straightforward.[18] Baba and Komatsuzaki suggested an approach (termed BK procedure hereafter) to extract free energy basins from the time series of a single distance.[19] The BK procedure is able to resolve different basins even if the distance distributions overlap because the short-time behavior of the observable is considered. Applied to a simplified model of a protein with 46 beads of three types (hydrophobic, hydrophilic, and neutral), the authors identified four free energy basins, in good agreement with the free energy surface derived using the complete structural information of the reference simulation.

Here we present a procedure for the automatic determination of free energy surfaces from single-molecule time series (FESST). First, an equilibrium transition network (ETN) is constructed by clustering individual time windows according to similarity in the short-time distribution of the signal, whose usage was inspired by the BK procedure.[19] The ETN is then used as the input for the minimum-cut-based free energy profile (cFEP) method, which is able to determine free energy basins and barrier heights (Figure 1).[7] The FESST parameters are optimized using an intrinsic cost function, the height of the unfolding barrier in the cFEP. This self-consistent choice of optimal FESST parameters leads to a unique solution in an objective and autonomous way, which allows for complete automatization of the procedure.

The accuracy of FESST is assessed using molecular dynamics (MD) trajectories of the 20-residue peptide β-3s,[20,21] whose sequence was designed to favor the three-stranded antiparallel β-sheet conformation, that is, a double β-hairpin.[22] β-3s has been shown to fold reversibly to the native structure determined by NMR[22] in MD simulations with the CHARMM polar hydrogen molecular mechanics potential energy function supplemented by a simple implicit solvent model.[23] In these simulations, β-3s folds in about 0.1 and 8 μs at 330 and 286 K, respectively.[24] Since multiple folding and unfolding events at the melting temperature of about 330 K can be simulated in less than a week (on a commodity processor), the free energy surface and the folding pathways and mechanism of β-3s have been

* To whom correspondence should be addressed. E-mail: schuler@bioc.uzh.ch; caflisch@bioc.uzh.ch.

[†] Current address: Empa, Swiss Federal Laboratories for Materials Science and Technology, Überlandstrasse 129, 8600 Dübendorf, Switzerland.

10.1021/jp1053698
Published on Web 10/22/2010

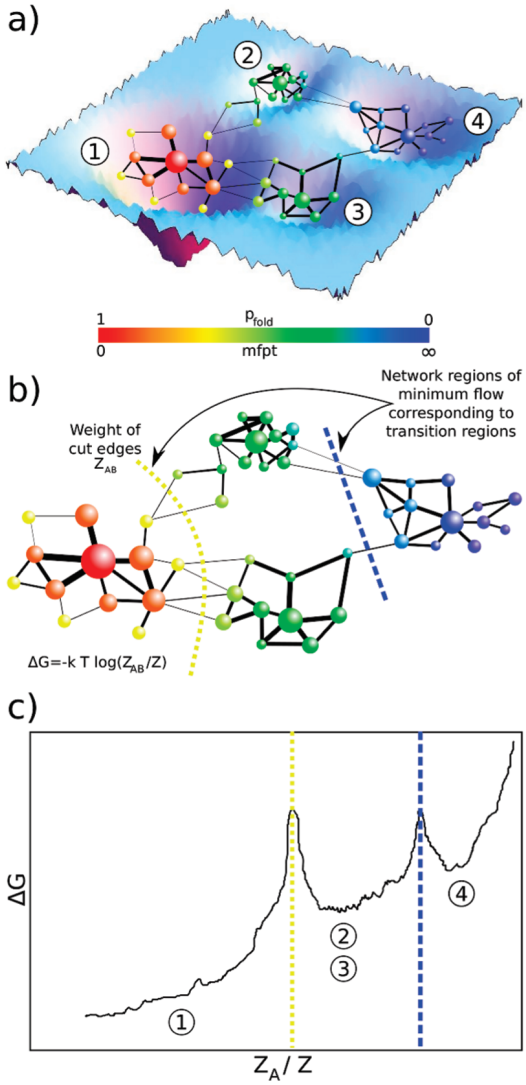

Figure 1. Illustration of the minimum-cut-based free energy profile (cFEP).[7] (a) The high-dimensional free energy surface is coarse-grained into nodes of the network. Two nodes are linked if the system proceeds from one to the other along the considered timeseries. The folding probability p_{fold} or the mean first passage time (mfpt) are calculated for each node analytically. Note that p_{fold} ranges from 1 (at the reference node) to 0 and mfpt from 0 to infinity. (b) For each value of p_{fold} (or mfpt), the set A of all nodes with a higher folding probability (or lower mfpt value) is defined. The free energy ΔG of the barrier between the two states formed by the nodes in A and the remainder of the network B can be calculated by the number of transitions Z_{AB} between nodes of either set.[7] (c) The cFEP is a projection of the free energy surface onto the relative partition function Z_A/Z, which includes all pathways to the reference node. For each value of p_{fold} (or mftp), the point (Z_A/Z, $-kT \log(Z_{AB}/Z)$) is added to the cFEP. The cFEP projects the free energy surface faithfully for all nodes to the left of the first barrier (basin 1). After the first barrier, two or more basins overlap (e.g., basins 2 and 3) if they have the same kinetic distance from the reference node.

investigated in detail.[6,20,21,25] The complexity of the free energy surface of β-3s[6] and its detailed characterization make it an ideal test system.

II. Methods

A. Free Energy Surface from Single-Molecule Time Series (FESST). FESST is a three-step procedure: construction of the ETN by clustering individual time windows using local kinetic information, identification of free energy basins by the cFEP approach, and removal of overlap from the non-native basins. The details of the three steps of FESST are presented in the next subsections and the Supporting Information (SI), while a schematic illustration is shown in Figure 2.

B. Coarse-Graining and Equilibrium Transition Network (ETN). Each time bin in the time series of the one-dimensional signal is assigned to a node of the ETN by the leader algorithm.[26] In the initialization step, the first bin is defined as the representative of the first node. At each successive bin t_n ($n > 1$), the distribution of the single-molecule observable within a short time window starting at time t_n (henceforth named the short-time distribution of t_n) is compared with the distributions of the previously visited representatives. The length of the time window is adjusted ideally such that it is about 1−5% of the characteristic time scale of the process monitored. To preserve the local kinetics (i.e., the actual dynamic evolution of the system), the comparison is carried out starting from the latest defined representative, that is, by parsing the list of representatives in inverse chronological order. A new node is defined whenever the short-time distribution of t_n deviates by more than a given threshold from the distributions of all previously defined representatives. In this way, one obtains a time series of nodes and a corresponding sequence of transitions between nodes, which is used to construct the ETN (Figure 2a,b).

C. Minimum-Cut-Based Free Energy Profile (cFEP). Krivov and Karplus have exploited an analogy between the kinetics of a complex process and equilibrium flow through a network to develop the cFEP, a projection of the free energy surface that preserves the barriers[7] and can be used for extracting folding pathways and mechanisms from MD simulations.[27] The input for the cFEP calculation is the ETN (Figure 1a), which is derived by the coarse-graining described above. For each node i in the ETN, the partition function is $Z_i = \sum_j c_{ij}$, that is, the number of times the node i is visited, where c_{ij} is the number of direct transitions from node i to node j observed along the time series. The transition probabilities can then be calculated as $p_{ij} = c_{ij}/\sum_k c_{ik}$. If the nodes of the ETN are partitioned into two groups A and B, where group A contains the reference node, then $Z_A = \sum_{i\in A} Z_i$ (the number of times a node in A is visited), $Z_B = \sum_{i\in B} Z_i$, and $Z_{AB} = \sum_{i\in A, j\in B} c_{ij}$ (the number of transitions between nodes in A and nodes in B). The free energy of the barrier between the two groups is $\Delta G = -kT \log(Z_{AB}/Z)$, where Z is the partition function of the full ETN (Figure 1b). The progress coordinate then is the normalized partition function Z_A/Z of the reactant region containing the reference node, but other progress coordinates can be used, because the cFEP is invariant with respect to arbitrary transformations of the reaction coordinate.[28]

In practice, the cFEP is calculated from the ETN in three steps: (1) The folding probability p_{fold} or the mean first passage time mfpt (Figure 1a) are calculated analytically for each node on the ETN by solving the system of transition rate equations.[7,27] (2) Nodes are sorted by decreasing values of p_{fold}, and for each of these values the relative partition function Z_A and the cut Z_{AB} are calculated (Figure 1b). (3) The individual points on the profile are evaluated as $[x = Z_A/Z, y = -kT \log(Z_{AB}/Z)]$ (Figure 1c). The result is a one-dimensional profile that preserves the barrier heights between the free energy basins; given the barriers, the basins can be determined.[7] It is important to note that the cFEP is a projection that preserves the heights of the barriers as long as the underlying coarse-graining does not group kinetically distant bins into the same node, that is, as long as the ETN captures the correct dynamics of the system. Time

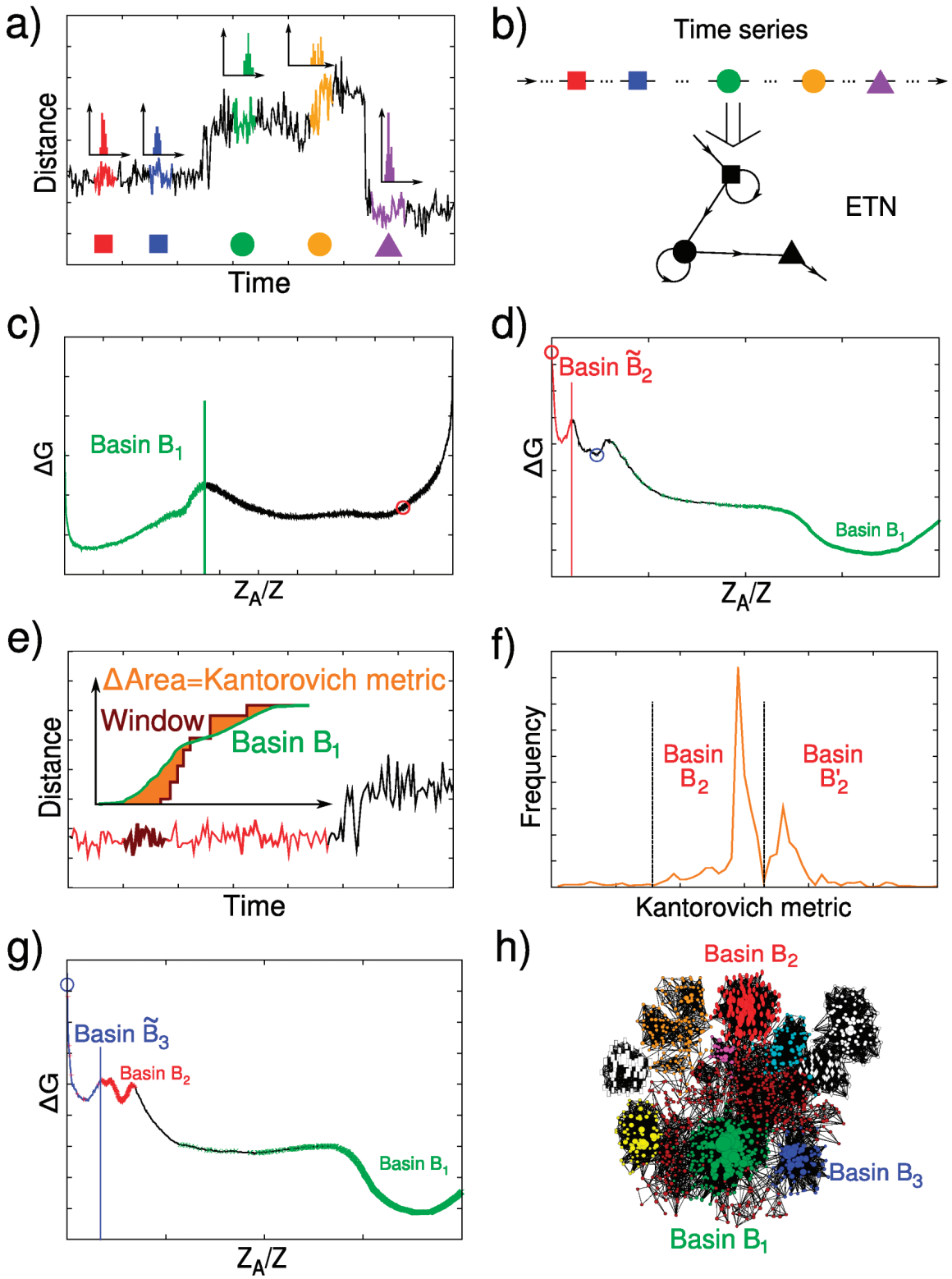

Figure 2. Schematic illustration of the FESST procedure. (a) The time series of the scalar signal is coarse-grained according to the short-time distribution of the distance. (b) The coarse-graining yields a time series of nodes and transitions that define the ETN. (c) The cFEP is plotted using the most populated node as a reference. The first free energy basin is isolated by cutting at the first barrier. The red circle indicates the most populated node outside the first basin, which is used to plot the cFEP for the determination of the second basin. (d) Because of the degeneracy of the short-term distance distribution, nodes from different free energy basins overlap in the second basin (see text). The tilde is used to denote a cFEP basin with overlap. The blue circle is the most populated node outside of $\tilde{B}_2$. (e,f) The overlap in $\tilde{B}_2$ is removed by comparing it with the entire distribution of the first basin (B_1). (g) The procedure is repeated for the next basin. (h) The basins extracted by FESST are illustrated on the conformation space network of β-3s with the native basin in green, and non-native basins Ch-curl$_1$ (B_2) and Ns-or$_1$ (B_3) in red and blue, respectively.[21]

bins misassigned by the coarse-graining result in spurious transitions between the basins and therefore lead to a lower barrier.[27]

All cFEPs in this paper are calculated with the software package WORDOM[29] using p_{fold} as the progress variable and an extra node for the $p_{fold} = 0$ boundary.[7]

D. Iterative Determination of Free Energy Basins. The most populated node is used to isolate the first basin by the cFEP approach. The barrier in the cFEP (Figure 2c) corresponds to the barrier leaving the basin identified first. For the remaining basins, the procedure is the same, except that the most populated unassigned node is used as a reference (Figure 2d). All nodes to the left of the cut at the first barrier make up the basin. Basins to the right of the first barrier are potentially overlapping (Figure 1c); thus, each basin requires a separate "exiting" profile.[27] Moreover, a FESST basin usually encompasses more than one of the true free energy basins, because the short-time distribution of the single distance can be degenerate. To remove this overlap, the signal's distribution in long time windows (ideally 10% of the time characteristic for the process monitored) starting from each bin assigned to the considered FESST basin is compared with the distribution of the signal in the entire basin identified previously (Figure 2e). Longer time windows are considered here for improved statistics and to exploit information complementary to the short-time distribution used in the construction of the ETN. Different subbasins in the basin to split are characterized by different ranges of the comparison metric (Figure 2f), because two distinct free energy basins differ in their similarity to a third one.

E. Static Model and Minimal-Kinetics Model. To investigate the importance of the system's kinetics and the signal's distribution in FESST, two simple procedures are tested for the identification of the native basin. They are called the static and the minimal-kinetics models as follows.

In the static model, a state is characterized by a range of observable values. To make the comparison with FESST as stringent as possible, the best possible static model is generated using the most accurate definition of the native basin.[27] For this purpose, the optimal observable range is determined for each completeness value (coverage of the native state by the basin identified, cf. Figure S3 of the SI) by testing multiple ranges and recording only the solution with the highest accuracy (fraction of the basin identified being native, cf. Figure S3).

In the minimal-kinetics model, each bin of the time series is assigned to the node defined by the discretized mean and standard deviation of the signal calculated over a short window around the bin considered. Subsequently, the resulting time series of nodes is analyzed by cFEPs as for FESST. This model incorporates local kinetic information by the consideration of the short-time evolution of the signal. In contrast to FESST, the minimal-kinetics model ignores the detailed structure of the signal's distribution. As for the static model, the parameters of the minimal-kinetics model, for example, the length of the window, are fine-tuned using an independent characterization of the native basin[27] as input.

III. Results. Application to MD Simulations of β-Sheet Folding

A. MD Simulations of β-3s. A total simulation time of 20 μs at 330 K was used for the FESST analysis. It has been shown previously that, in MD at 330 K, β-3s folds reversibly to the NMR conformation, irrespective of the starting structure; 23 of the 26 nuclear Overhauser effect constraints are satisfied.[20,21] All MD runs and most of the analysis of the trajectories were carried out with CHARMM.[30]

A mean field approximation based on the solvent accessible surface (SAS) was used to describe the main effects of the aqueous solvent.[23]

B. Intramolecular Distance and Metric Used for Coarse-Graining. The time series of the $C_\beta Gln_4 - C_\beta Thr_{16}$ distance is used in FESST, but the results are robust with respect to the choice of residue pairs as long as one of the two residues is in β-strand 1 and the other in β-strand 3. Two time windows $[t_1, t_1 + \tau]$ and $[t_2, t_2 + \tau]$ are grouped into the same node of the ETN if their distributions of the intramolecular distance (the short-term distribution) pass a Kolmogorov−Smirnov test,[31] which checks if two samples are picked from the same distribution. Each MD snapshot is used as a starting point of a time series bin, so that there are as many bins as coordinate frames along the MD trajectory. In other words, two successive bins are shifted by the MD saving interval of 20 ps. The length of the time window τ is chosen such that it is much shorter than the folding time, which is about 100 ns in MD simulations of β-3s at 330 K.[21] The dissimilarity of the two time windows is defined as the maximum difference of the cumulative distribution functions c_1, c_2 of the distance r (dissimilarity $= \max_{r>0}|c_1(r) - c_2(r)|$). The test is passed; that is, two time bins are grouped in the same node if

$$\text{dissimilarity} \leq \sqrt{\frac{2}{N}} \cdot \zeta$$

where N is the number of MD snapshots in each time window and ζ the acceptance cutoff that corresponds to a certain confidence level.[31] Note that the FESST results on β-3s are robust with respect to the choice of N in the range $30 \leq N \leq 250$ (i.e., 0.6 ns $\leq \tau \leq$ 5 ns) and ζ in the range $0.3 \leq \zeta \leq 1.5$ (Figure S1 of the SI). Values of $\tau = 2$ ns and $\zeta = 0.3$ are used in the following. To slightly improve on the sampling of the MD simulation, a detailed balance is imposed to the FESST-ETN by averaging the numbers of transitions c_{ij} and c_{ji} between nodes i and j.

C. Native Basin and Unfolding Barrier. The free energy basins of β-3s have been determined previously by the cFEP procedure using information on all 645 coordinates.[27] Since the full information of the peptide dynamics was taken into account, those free energy basins and barriers are used here as a reference for a critical evaluation of FESST and the comparison with other approaches.

Using the time series of the $C_\beta Gln_4 - C_\beta Thr_{16}$ distance, the native basin of β-3s is determined by FESST with remarkable accuracy (96% of the FESST native basin is part of the native state as determined by the 645 coordinates of cFEP, cf. Figure S3 of the SI) and completeness (95% of the native state of the 645 coordinates of the cFEP is captured by the FESST native basin, cf. Figure S3 of the SI). Moreover, the FESST unfolding barrier (defined as the free energy difference between the bottom of the first basin on the left in the cFEP and the top of the first barrier in the cFEP[7]) has a height (10.7 kJ/mol) very similar to the one obtained by the 645-coordinates cFEP (10.6 kJ/mol, Figure 3a and Figure S5 of the SI).

To investigate the influence of the choice of the residue pair monitored, each of the 154 $C_\beta - C_\beta$ pairs was tested in FESST. Remarkably, for 32 of these pairs the native basin is identified with an accuracy greater than 80% and at the same time a completeness of more than 90% (Figure 4b). Interestingly, the larger the separation along the sequence, the better the score. A notable exception is the 5−7 distance, which reflects the formation of the β-turn at the N-terminal hairpin. The distances yielding the best score are those between residues in β-strands 1 and 3 (top left part of the matrix in Figure 4b), which is likely to be a consequence of the β-sheet topology. Moreover, the $C_\beta - C_\beta$ distances involving the N-terminal β-strand show a higher score than those involving the C-terminal β-strand, which is consistent with the higher structural stability of the C-terminal

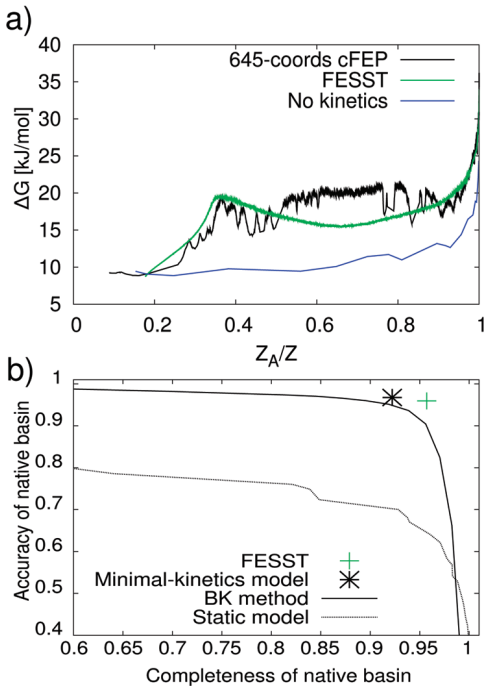

Figure 3. (a) Determination of the native basin by taking into account all structural information (black) or only a single distance (green). The cFEP is shown with the most populated node as a reference (for details, see the SI). If the short-time kinetics of the system is ignored, (only the network of transitions between the coarse-grained values of the single distance is analyzed), the cFEP (blue) displays no discernible barrier, and no meaningful basin can be extracted. (b) Comparison of FESST with a previously published single distance approach (BK = Baba and Komatsuzaki[19]) and two simple models (see the Model section for details). Note that FESST and the minimal-kinetics model yield a single data point rather than an accuracy versus completeness curve, because there is a unique optimum that can be determined by maximizing the barrier height.

hairpin.[20,21] In other words, the fully folded state can be better separated from non-native conformers by taking into account the N-terminal β-strand, because the C-terminal hairpin is folded correctly in the most populated non-native conformers.

FESST performs much better than the static model (Figure 3b and Figure S8 of the SI), which shows the importance of exploiting information about the short time kinetics. The cFEP calculated from the time series of coarse-grained distance values without taking into account the kinetic information displays no barrier, which indicates that a single distance value does not discriminate the native state of β-3s (Figure 3). If, in FESST, the time series was coarse-grained based on statistical properties (such as mean and standard deviation) instead of the signal's distribution, a native state of comparable quality would result (minimal-kinetics model in Figure 3b), but non-native basins are detected significantly less accurately (e.g., only 72% accuracy and 77% completeness for Ch-curl$_1$, which is defined below).

Although both approaches make use of short-time distributions of the signal, FESST has two advantages compared to the BK procedure.[19] First, FESST exploits the local kinetic information for the coarse-graining, while the BK procedure iteratively removes the time windows least similar to the distribution of the whole distance time series, thus ignoring the chronological order of the windows. Second, the optimal values of parameters required by FESST (size of the time window k and acceptance cutoff ζ used in the coarse-graining) can be determined automatically using the cFEP barrier height as a cost function, because the barrier height is the main determinant of the interconversion rates between the free energy basins. Therefore, the most accurate determination of the native basin yields the highest barrier.[27] Correspondingly, the parameter set yielding the highest barrier achieves the highest score (defined as the product of accuracy and completeness, Figure S2 of the SI). Therefore, FESST yields a single data point in the accuracy versus completeness plot (Figure 3b), whereas the basins extracted by other procedures depend on the cutoffs chosen for their iterative refinement, so that it is not possible to automatically identify the optimal solution.

D. Identification of Non-native Basins. The most populated node outside of the native basin is used as a reference to plot the cFEP profile for identifying the first non-native basin (termed $\tilde{B}_2$ in Figure 2). Because of the degeneracy of the short-time distribution of the distance, multiple free energy basins may overlap on the cFEP. Such overlap can be removed by comparing the long-time distance distribution of each time window with the distance distribution in a previously identified basin. In practice, for each time window $[t_2, t_2 + T]$ in basin $\tilde{B}_2$, the distribution of the distance is compared with the histogram of the entire native basin. Time windows of length

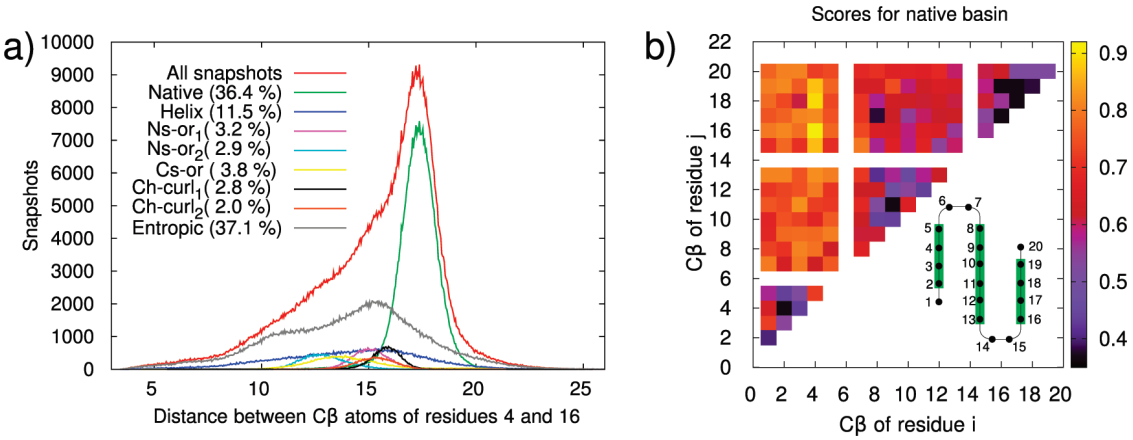

Figure 4. Robustness of FESST with respect to the choice of the distance. (a) Histograms of distance between the C_β atoms of residues 4 and 16 for the snapshots in each free energy basin determined by cFEP using all 645 degrees of freedom of β-3s.[27] (b) Matrix of scores for native state detection. Each (i,j) value of the score was calculated by applying FESST to the time series of distance between the C_β-atoms of residues i and j (Gly_6 and Gly_{14} have no C_β atom). The inset shows a schematic representation of β-3s with the three native β-strands (green rectangles). In the coarse-graining step of FESST, $N = 100$ distance values are compared, and the acceptance cutoff $\zeta = 0.3$ is used in the Kolmogorov−Smirnov test (cf. Section III B).

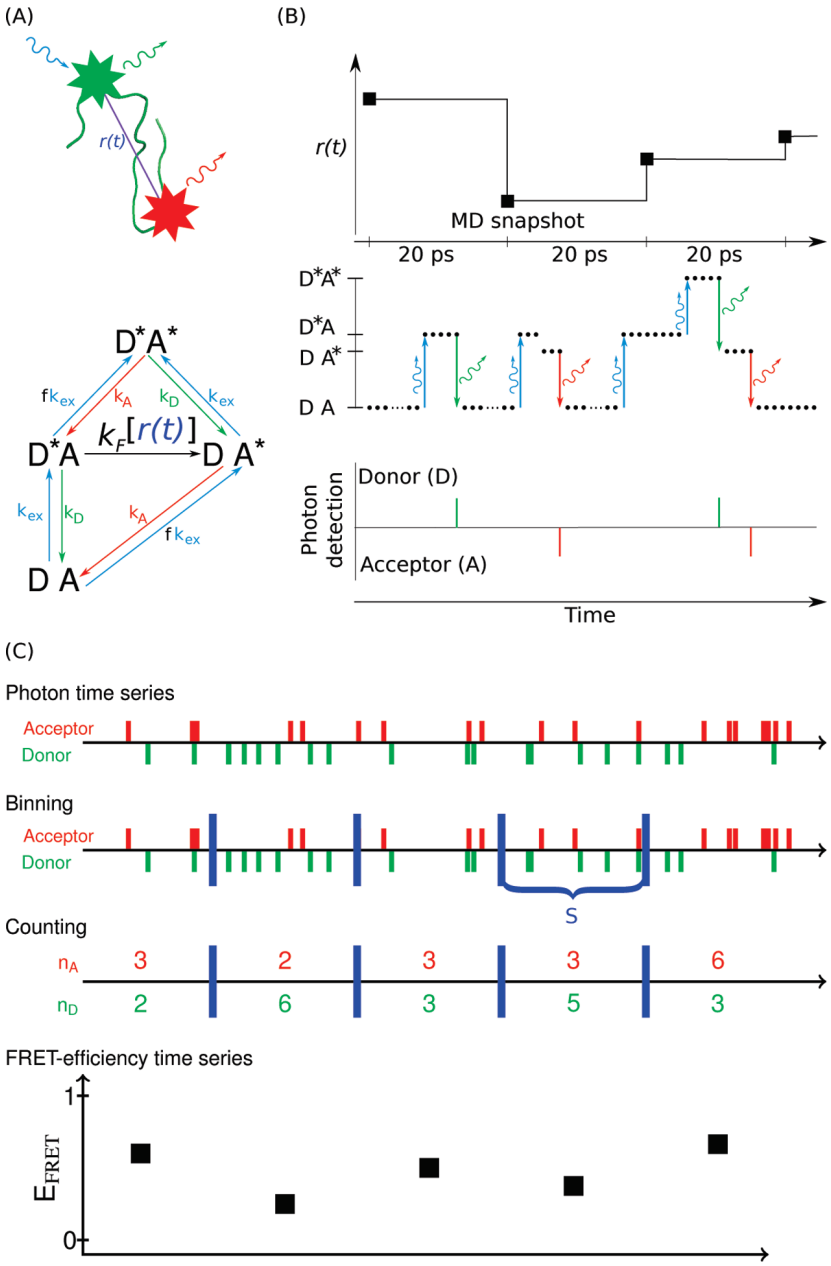

Figure 5. Markov state model used to generate the photon time series from the MD time series and illustration of the transformation of the photon time series into the FRET-efficiency time series. Red, green, and blue curled arrows represent acceptor photon emission, donor photon emission, and photon absorbance. (A, top) Schematic illustration of the emulated FRET experiment, where $r(t)$ represents the distance between the C_β-atoms of residues 4 and 16. (A, bottom) State diagram of the Markov process used to simulate the FRET experiment. (B) Illustration of the simulation of the FRET experiment, which uses the time series of the distance r, measured along the MD trajectory, to generate the photon time series. For each MD saving interval of 20 ps, 100 steps (black dots) of a random walker on the Markov state model are carried out with a constant value of the distance $r(t)$, i.e., the constant Förster rate $k_F[r(t)]$. Note that each excitation leads to an emitted photon in the FRET emulation. Direct excitation of the acceptor is set to $f = 5\%$ of the donor excitation rate.[40] (C) Transformation of the photon time series into the FRET-efficiency time series. The initial photon time series is binned with binning time S. In each bin, the number of acceptor photons n_A and donor photons n_D is counted. The FRET efficiency is then calculated using the formula $E_{FRET} = n_A/(n_A + n_D)$.

$T \approx (10 \text{ to } 20)\ \tau$, that is, significantly larger than those used for the construction of the ETN, are considered here for better statistics. The comparison consists of calculating the Kantorovich metric[32] between the two distributions (the area between the two cumulative histograms, Figure 2e). Finally, each peak in the histogram of the Kantorovich values is assigned to a new subbasin (Figure 2f). The window size T can be chosen by optimizing the separation of the different peaks in the Kantorovich histogram (Figure S9 of the SI).

With this procedure, the basin B_2 derived from $\tilde{B}_2$ corresponds to the 645 coordinates of the free energy basin Ch-curl$_1$ (curl-like conformation with folded C-terminal hairpin[27]) with 92% accuracy and 85% completeness. Further, the third FESST basin $\tilde{B}_3$ encompasses two free energy basins and can be split by comparing with the distance distribution in $\tilde{B}_2$. The free energy basin Ns-or$_1$ (N-terminal strand out of register and folded C-terminal hairpin[27]) can be extracted with 77% accuracy and 68% completeness. The second subbasin detected in B_3 contains 56% of MD snapshots in Ch-curl$_2$ (curl-like conformation 2[27]) covering 77% of these MD snapshots. These non-native conformers are stabilized mainly enthalpically.[27] Entropically stabilized conformations such as those in the "helical basin" and the "entropic region"[27] show a very broad distribution of distances (blue and gray curves in Figure 4a). These broad

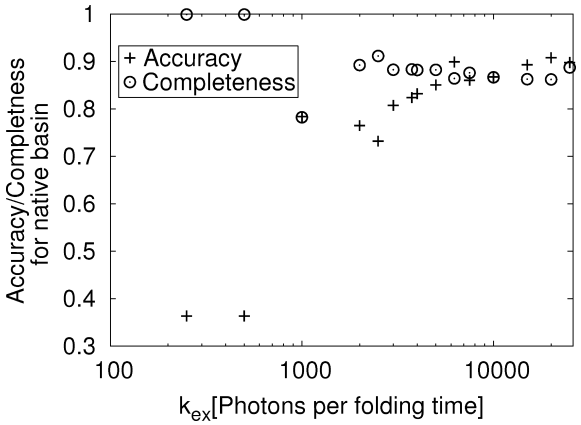

Figure 6. FESST performance on an emulated FRET experiment. The time series of FRET efficiencies calculated for 0.4 ns bins is used for the analysis by FESST with a window size of 25 bins (the effect of other window sizes is illustrated in Figure S11 of the SI) and acceptance cutoff $\zeta = 0.3$. The accuracy and completeness for the identification of the native basin is calculated for the set of snapshots in all bins of the first FESST basin (cf. Figure S3 for the definition of accuracy and completeness). The excitation rate k_{ex} is expressed as the average number of photons per folding time (about 100 ns for β-3s[21] in MD simulations at 330 K).

distributions overlap strongly with those of other basins and therefore are distributed over multiple FESST basins. In other words, both $\tilde{B}_2$ and $\tilde{B}_3$ contain time windows of the entropic region that can be removed by the procedure illustrated in Figure 2e,f. (For the native basin this step is not performed because the overlap of the distance distributions is much smaller than for $\tilde{B}_2$ and $\tilde{B}_3$, and no basin for comparison is available.)

IV. Application to an Emulated FRET Signal

A promising experimental method for obtaining intramolecular distance information in heterogeneous systems is single-molecule FRET.[12,15,17] To elucidate the applicability of FESST to such data, a FRET experiment is mimicked by generating a photon time series using a Markov state model (Figure 5). In this model, the rate of energy transfer $k_F(r)$ between the two "virtual" chromophores depends on the inverse sixth power of the distance r between the C_β-atoms of β-3s residues 4 and 16 as recorded along the MD trajectory (for details, see SI). From such photon time series, the native state of β-3s can be detected with 78% accuracy and 78% completeness from 1000 photons per folding time (Figure 6). This detection quality is obtained by comparing intervals of the time series of FRET efficiencies (for details, cf. Figure 5C and SI) as long as 10 ns, which corresponds to about one tenth of the folding time of β-3s in the MD simulation at the melting temperature. The detection quality depends only weakly on the size of each FRET bin (Figure S10 of the SI) and the length of the time series interval (Figure S11 of the SI). However, simulations of a simple two-state model indicate that the required number of photons is significantly reduced if the individual populations are slightly better separated in distance space (details in SI) than the free energy basins of β-3s (Figure 4a), suggesting that the application of the method to experiments is feasible.

V. Application to Single Molecule FRET Experiments

In a first attempt to apply FESST to real experimental data, we chose a protein whose folding dynamics at the unfolding midpoint (where both folded and unfolded state are populated) are in the range of a few milliseconds. This allows us to use experiments on freely diffusing molecules to maximize the excitation rate, while there is at the same time a large probability of observing folding or unfolding transitions during the diffusion time through the confocal volume of about a millisecond. As a result, we obtain a large number of short observations, similar in spirit to short simulations from parallel or distributed computing.[33] Even though these short observations will largely be independent, they can still be used to reconstruct the free energy surface if they are locally equilibrated, representative of the relevant conformational space, and provide a sufficient time resolution for the process investigated.[24,33]

We used a variant of monomeric λ-repressor, a well-established fast-folding protein[34,35] with a folding relaxation time of about 1 ms at the unfolding midpoint,[36] labeled it with Alexa Fluors 488 and 594 as FRET donor and acceptor, respectively, and investigated it with confocal single molecule experiments at a guanidinium chloride concentration of 0.68 M (see SI for details). Every protein molecule diffusing through the confocal volume will then result in a burst of photons. By selecting bursts with a duration of at least 2.5 ms and by working at high excitation rate, the average number of photons in the 48461 events collected during 4 h measurement time used was 575, which allows us to apply the FESST analysis with a data binning of 0.1 ms. FESST correctly identifies the folded and unfolded subpopulations (Figure 7a). Because of the high excitation power used, a significant contribution of acceptor photobleaching is present, which results in a third apparent population at lower transfer efficiencies.

The corresponding cFEP plotted from the unfolded state (Figure 7b) exhibits a well-defined folding barrier at $Z_A/Z = 0.42$, whose height can be maximized by node merging as for the analysis of the β-3s simulations (details in the SI, Figure S5). Even though the position of the barrier agrees with the relative populations, the resulting barrier height for folding (and for unfolding, see Figure S16 of the SI) is in the range of 1 kT, significantly lower than expected from the folding rate of 225 s^{-1} extracted from the frequency of transitions identified in the bursts. A factor that contributes to the reduction in barrier height is the imperfect separation of populations due to shot noise, resulting in spurious transitions in the FESST analysis. Another limitation is the time resolution achievable with photon detection rates in the range of about 0.14 MHz currently available in our free diffusion experiments, with which it is not possible to resolve the nanosecond diffusive dynamics of the polypeptide chain[16] from individual fluorescence bursts. Even though FESST is still limited by the photon rates in current single molecule experiments, the clear identification of subpopulations and the existence of a barrier in the resulting free energy profile illustrate its feasibility and potential for the analysis of experimental data.

VI. Discussion

FESST is a method for determining free energy basins and barriers from the time evolution of a scalar observable. The accuracy and range of possible applications of FESST have been investigated using the scalar time series derived from atomistic MD simulations of the reversible folding of a structured peptide. First, FESST was applied to the time series of a single interresidue distance of β-3s, a 20-residue peptide with native three-stranded β-sheet topology. The native state of β-3s, three subbasins in the denatured state, and the free energy barrier for unfolding can be determined with high accuracy. Importantly, FESST is robust to the choice of the residue pair. In fact, 20% of the 154 pairs of $C_\beta - C_\beta$ distances can be used in FESST for

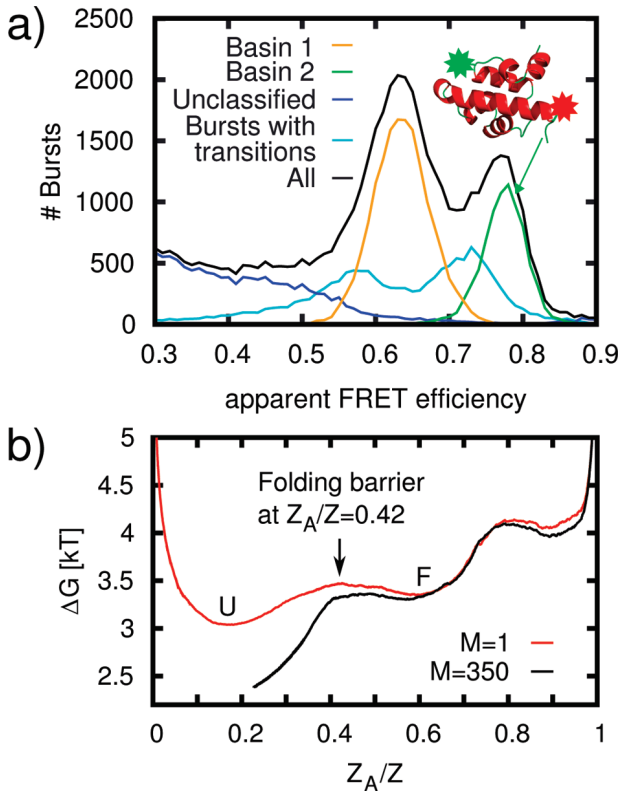

Figure 7. FESST analysis of single-molecule FRET measurements on a monomeric λ-repressor. (a) Histogram of apparent FRET efficiency of all identified bursts (black). The histograms of the apparent FRET efficiency for bursts containing only bins attributed by FESST to the first (unfolded) and the second (folded) basin are shown in orange and green, respectively. Those bursts not assigned to basins 1 or 2 are shown in blue. Because of averaging, the apparent FRET efficiency of the bursts with transitions between the FESST basins cumulates between the mean FRET efficiency of the other FESST basins (cyan curve). The inset shows the structure of the folded λ-repressor fragment with the FRET labeling sites. For the FESST coarse-graining, a window size $N = 25$ and an acceptance cutoff $\zeta = 0.4$ are used (the effect of other coarse-graining parameters is shown in Figure S15 of the SI). (b) One-dimensional projection of the free energy surface of λ-repressor. The cFEP is plotted using the most populated node in the FESST-ETN as the reference, in this case a representative of the unfolded basin (U). The location of the barrier for folding is indicated (black arrow). The highest barrier for folding is found when $M = 350$ nodes in the U basin are merged (black). The second barrier at $Z_A/Z = 0.8$ originates from the transitions to the bleached state.

determining the native state of β-3s, and in particular distances between residues in β-strands 1 and 3 are optimal. Furthermore, the basin assignment by FESST is robust to changes of the parameters used for coarse-graining, which can be determined self-consistently.

In a second test, FESST was applied to a time series of FRET efficiencies generated from the MD trajectory. An accurate identification of the native basin of β-3s is possible with FRET efficiencies calculated from about 1000 photons emitted during the folding time.

A first application to single-molecule FRET experiments on a freely diffusing monomeric λ-repressor with folding dynamics in the millisecond range shows that FESST is able to correctly identify the folded and unfolded subpopulations and yields a free energy profile that captures this separation. This result clearly demonstrates the feasibility of applying FESST to experimental data. However, the height of the folding barrier in the corresponding free energy profile is lower than expected, an effect that is presumably dominated by the current limitations in photon rates. Recent developments in the use of additives that reduce photobleaching and increase fluorescence emission rates[37−39] are expected to contribute strongly to an improvement of this situation both for experiments on freely diffusing and immobilized molecules.

The present analysis focused on the FRET efficiency, because it is one of the most commonly used observables. Additional information, for example, interphoton times, polarization, or fluorescence lifetimes, is expected to further increase the discriminatory power of FESST. In conclusion, FESST can be applied to the time series of any type of scalar observable as long as the short-time distribution of the single-molecule signal contains enough information to allow FESST to remove the signal's degeneracy.

Acknowledgment. We thank Dr. D. Nettels for help with the data analysis and Drs. S. Muff, I. Gopich, D. Nettels, and S. V. Krivov for interesting discussions. This work was supported by grants of the Swiss National Science Foundation to A.C. and B.S. and a Starting Investigator Grant of the European Research Council (FP7) to B.S. Most of the simulations were carried out on the Matterhorn computer cluster and on the Schrödinger computer cluster of the University of Zurich.

Supporting Information Available: FESST applications to atomistic simulations of β-sheet folding, to an emulated FRET signal, and to real experimental data and the FESST resolution limit. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Goldstein, M. *J. Chem. Phys.* **1969**, *51*, 3728–3739.
(2) Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. *Science* **1991**, *254*, 1598–1603.
(3) Dill, K. A.; Chan, H. S. *Nat. Struct. Biol.* **1997**, *4*, 10–19.
(4) Shakhnovich, E. I. *Phys. Rev. Lett.* **1994**, *72*, 3907–3910.
(5) Krivov, S. V.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 14766–14770.
(6) Rao, F.; Caflisch, A. *J. Mol. Biol.* **2004**, *342*, 299–306.
(7) Krivov, S. V.; Karplus, M. *J. Phys. Chem. B* **2006**, *110*, 12689–12698.
(8) Caflisch, A. *Curr. Opin. Struct. Biol.* **2006**, *16*, 71–78.
(9) Noé, F.; Fischer, S. *Curr. Opin. Struct. Biol.* **2008**, *18*, 154–162.
(10) Berezhkovskii, A.; Hummer, G.; Szabo, A. *J. Chem. Phys.* **2009**, *130*, 205102.
(11) Bai, C.; Wang, C.; Xie, X. S.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 11075–11076.
(12) Ha, T.; Enderle, T.; Ogletree, D. F.; Chemla, D. S.; Selvin, P. R.; Weiss, S. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 6264–6268.
(13) Deniz, A. A.; Laurence, T. A.; Beligere, G. S.; Dahan, M.; Martin, A. B.; Chemla, D. S.; Dawson, P. E.; Schultz, P. G.; Weiss, S. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 5179–5184.
(14) Schuler, B.; Lipman, E. A.; Eaton, W. A. *Nature (London)* **2002**, *419*, 743–747.
(15) Haran, G. *J. Phys.: Condens. Matter* **2003**, *15*, R1291–R1317.
(16) Nettels, D.; Gopich, I. V.; Hoffmann, A.; Schuler, B. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 2655–2660.
(17) Schuler, B.; Eaton, W. A. *Curr. Opin. Struct. Biol.* **2008**, *18*, 16–26.
(18) Li, C.-B.; Yang, H.; Komatsuzaki, T. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 536–541.
(19) Baba, A.; Komatsuzaki, T. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 19297–19302.
(20) Ferrara, P.; Caflisch, A. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 10780–10785.
(21) Muff, S.; Caflisch, A. *Proteins: Struct., Funct., Bioinf.* **2008**, *70*, 1185–1195.
(22) De Alba, E.; Santoro, J.; Rico, M.; Jiménez, M. A. *Protein Sci.* **1999**, *8*, 854–865.
(23) Ferrara, P.; Apostolakis, J.; Caflisch, A. *Proteins: Struct., Funct., Bioinf.* **2002**, *46*, 24–33.
(24) Muff, S.; Caflisch, A. *J. Phys. Chem. B* **2009**, *113*, 3218–3226.
(25) Cavalli, A.; Haberthür, U.; Paci, E.; Caflisch, A. *Protein Sci.* **2003**, *12*, 1801–1803.

(26) Hartigan, J. *Clustering Algorithms*; Wiley: New York, 1975.
(27) Krivov, S. V.; Muff, S.; Caflisch, A.; Karplus, M. *J. Phys. Chem. B* **2008**, *112*, 8701–8714.
(28) Krivov, S. V.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 13841–13846.
(29) Seeber, M.; Cecchini, M.; Rao, F.; Settanni, G.; Caflisch, A. *Bioinformatics* **2007**, *23*, 2625–2627.
(30) Brooks, B. R.; Brooks, C. L., III; Mackerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; et al. *J. Comput. Chem.* **2009**, *30*, 1545–1614.
(31) Smirnov, N. V. *Mat. Sb.* **1939**, *6*, 3–24.
(32) Vershik, A. *J. Math. Sci.* **2006**, *133*, 1410–1417.
(33) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 19011–19016.
(34) Huang, G. S.; Oas, T. G. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 6878–6882.
(35) Yang, W. Y.; Gruebele, M. *Nature (London)* **2003**, *423*, 193–197.
(36) Ghaemmaghami, S.; Word, J. M.; Burton, R. E.; Richardson, J. S.; Oas, T. G. *Biochemistry* **1998**, *37*, 9179–9185.
(37) Rasnik, I.; McKinney, S. A.; Ha, T. *Nat. Methods* **2006**, *3*, 891–893.
(38) Widengren, J.; Chmyrov, A.; Eggeling, C.; Löfdahl, P.-A.; Seidel, C. A. M. *J. Phys. Chem. A* **2007**, *111*, 429–440.
(39) Vogelsang, J.; Kasper, R.; Steinhauer, C.; Person, B.; Heilemann, M.; Sauer, M.; Tinnefeld, P. *Angew. Chem., Int. Ed.* **2008**, *47*, 5465–5469.
(40) Schuler, B. *Methods Mol. Biol.* **2007**, *350*, 115–138.

JP1053698

Supporting Information

Contents

List of Figures

1 FESST application to atomistic simulations of β-sheet folding

Most of the data presented in the first subsection of this supporting information (SI) refer to the time series of the single distance C_β Gln_4−C_β Thr_{16} in the Beta3s peptide whose time series was generated by atomistic simulations (cf. Sec. 1.1). Robustness tests are presented in subsections 1.2-1.5.

1.1 Molecular dynamics (MD) simulations

Beta3s is a designed 20-residue peptide (Thr_1-Trp_2-Ile_3-Gln_4-Asn_5-Gly_6-Ser_7-Thr_8-Lys_9-Trp_{10}-Tyr_{11}-Gln_{12}-Asn_{13}-Gly_{14}-Ser_{15}-Thr_{16}-Lys_{17}-Ile_{18}-Tyr_{19}-Thr_{20}) that folds to a three-stranded anti-parallel β-sheet [1, 2]. Multiple folding and unfolding events have been sampled by molecular dynamics (MD) simulations [3, 4] with an implicit solvent model [5]. In these MD simulations, Beta3s was modeled by explicitly considering all heavy atoms and the hydrogen atoms bound to nitrogen or oxygen atoms (PARAM19 force field [3, 6] with the default cutoff of 7.5 Å for the nonbonding interactions). A mean field approximation based on the solvent accessible surface (SAS) was used to describe the main effects of the aqueous solvent [5]. More explicitly, the screening of the electrostatic interactions is approximated by the distance-dependent dielectric function $\epsilon(r) = 2r$, while the remaining solvation effects are approximated by replacement of the monopole moment of charged groups by strong dipole moments and a linear function of atomic SAS values. The latter requires only two surface-tension like parameters and takes into account both polar and apolar solvation effects by a negative (i.e., favorable) value of the surface-tension parameter for nitrogen and oxygen atoms, and a positive (unfavorable) value for carbon and sulfur atoms. Ten MD runs of 2 μs each with different initial distributions of velocities were performed with the Berendsen thermostat (coupling constant of 5 ps) at 330 K, which is slightly above the melting temperature of Beta3s [7]. A time step of 2 fs was used and the coordinates were saved every 20 ps for a total of 10^6 MD snapshots. This required three weeks on a 10-CPU cluster.

1.2 Robustness of FESST upon variation of window size and cutoff for coarse-graining

It is important to evaluate the performance of FESST upon changes in the parameters used for coarse-graining. This analysis shows that the determination of the native basin is robust (Fig. S1). The height of the unfolding barrier as determined in the cFEP can be used to find the optimal coarse-graining parameters (Fig. S2).

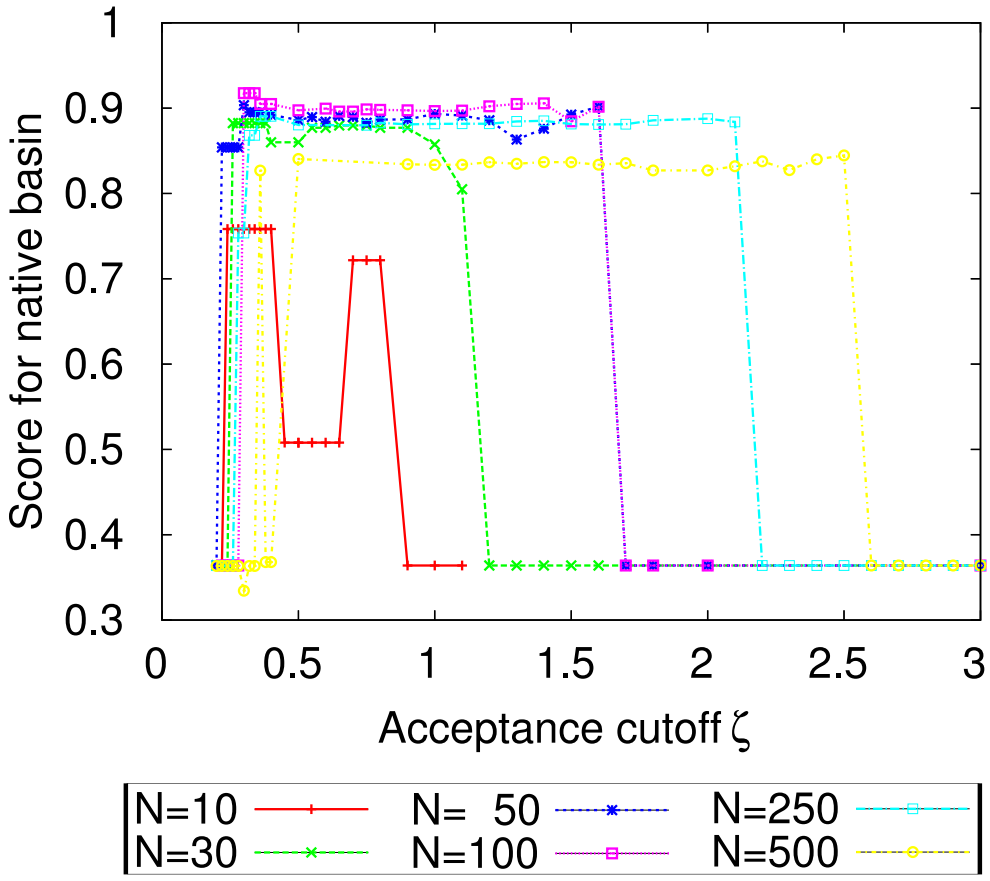

Figure S1: Robustness of FESST upon variation of the parameter used for coarse-graining. The signal is the single distance C_β Gln_4–C_β Thr_{16} of Beta3s obtained by implicit solvent MD. The score is the product of accuracy (i.e., fraction of the FESST native basin belonging to the native state as determined using all 645 coordinates of Beta3s, cf. Fig. S3) and completeness (i.e., fraction of the native state captured by FESST, cf. Fig. S3). The range of values tested for the size of the time window is $10 \leq N \leq 500$, i.e., 0.2 ns $\leq \tau \leq$ 10 ns as the number of MD snapshots N is equal to τ times the saving frequency of $1/20$ ps^{-1}. Results in the main text are obtained for values of $\tau = 2$ns (N=100) and $\zeta = 0.3$.

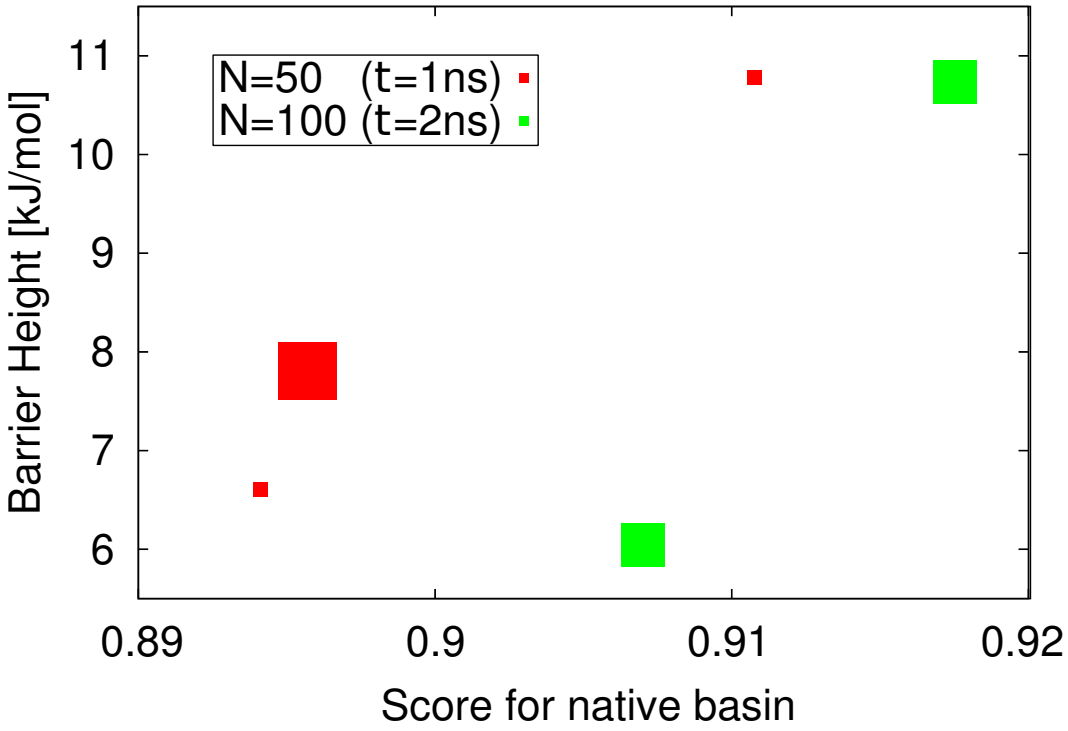

Figure S2: Effect of the parameters used for coarse-graining on the FESST determination of the native state on the height of the unfolding barrier in the cFEP. Note that multiple values of the threshold ζ yield the same score and barrier height. The size of the symbol is proportional to the number of values tested. As an example, the best result, i.e., the data point with highest score *and* barrier height (green square in the top right corner) is obtained with $\zeta = 0.30, \zeta = 0.32,$ and $\zeta = 0.34$ using a window size of N=100. The plot provides evidence that the parameters can be optimized with the height of the cFEP barrier as a cost function.

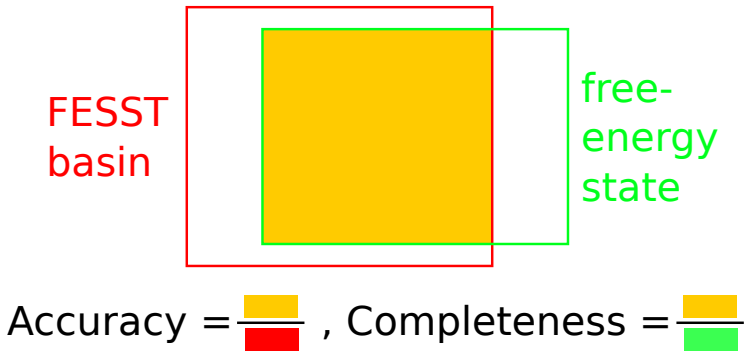

Figure S3: Definition of accuracy and completeness for the free-energy state as determined using the information of all 645 coordinates of Beta3s. The accuracy is the fraction of the FESST basin belonging to the free-energy state as determined using all 645 coordinates of Beta3s, and the completeness is the fraction of the free-energy state captured by FESST.

1.3 Merging of nodes in the native basin

In comparison to the 645-coordinates equilibrium transition network (ETN), the ETN derived from the single distance signal (termed FESST-ETN) lacks nodes with very large weight (Fig. S4). For instance, the most populated node in the latter (158 snapshots) is almost three orders of magnitude smaller than the most visited node of the 645-coordinate ETN (88022 snapshots). To render the most populated node in the FESST-ETN more representative of the native basin, the M heaviest native nodes are combined. The native basin consists of those nodes with a value of the progress variable Z_A/Z in the cFEP (calculated from the most populated node) smaller than the value at the first barrier in the cFEP[8]. The new ETN is constructed from the node sequence in the MD simulation with the heaviest M native nodes merged into one node. The merging step affects only the native basin (inset of Fig. S5), and mainly results in a lower (i.e., more favorable) free-energy value for the bottom of the native basin, i.e., a higher unfolding barrier. The highest value of the unfolding barrier is observed for $M = 7000$ nodes merged. The value of the barrier height is robust for $3000 \leq M \leq 12000$. For $M \geq 3000$ nodes merged, the weight of the most populated node in the FESST-ETN exceeds the weight of the most visited node in the unmodified 645-coordinates ETN (Fig. S6). As a basis of comparison, the merging procedure can also be applied to the 645-coordinates ETN. The highest barrier is found for only 107 nodes merged and exceeds the value for the unmodified network by only 0.7 kJ/mol (Fig. S5).

Another consequence of the reduced size of the reference node is the overestimation of the time needed to reach the reference node from the other nodes in the FESST-ETN, i.e., a folding time much longer than the one obtained from the 645-coordinates ETN (Fig. S7). For the FESST-ETNs with $M \geq 3000$, the distribution of the mean first passage times matches those of the 645-coordinates ETN (Fig. S7). The correspondence of the folding time distributions provides further evidence that the system dynamics is reliably captured by the FESST coarse-graining. The reliable representation of the system's dynamics in the ETN is a necessary condition for the correct operation of the cFEP approach [8].

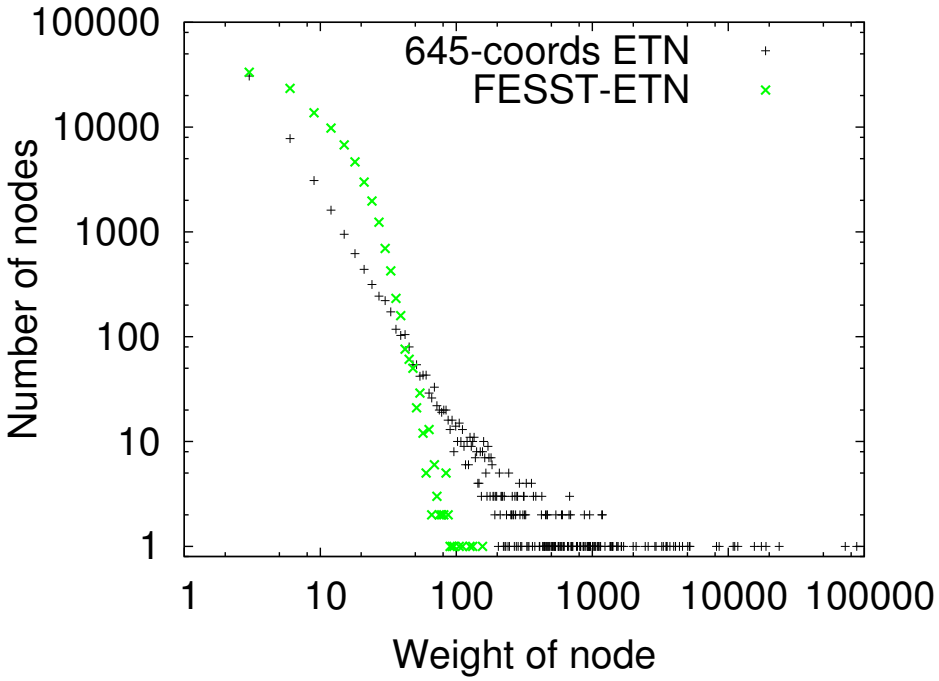

Figure S4: Distribution of node weights (number of snapshots) for the 645-coords ETN and the FESST-ETN. The distance C_β Gln_4−C_β Thr_{16} in Beta3s, window size N = 100, and acceptance cutoff $\zeta = 0.3$ are used. Note that the weight of the most populated node is 88022 and 158 for the 645-coords ETN and the FESST-ETN, respectively.

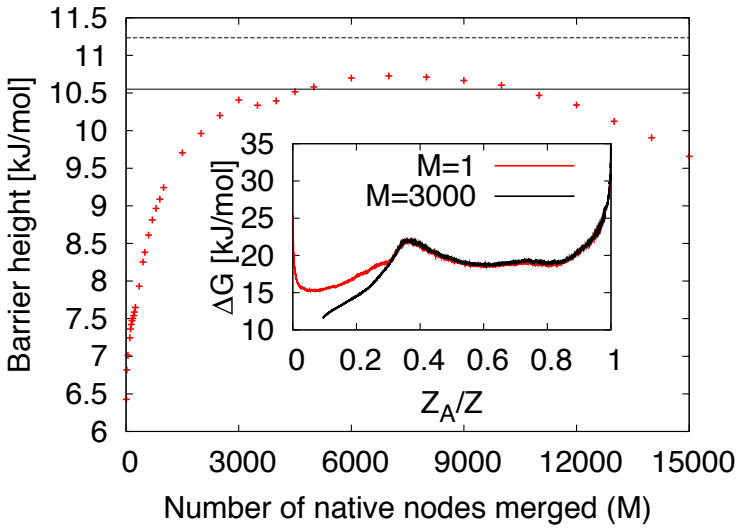

Figure S5: Dependence of barrier height on the merging of the heaviest nodes of the native basin. The red crosses show the barrier height of the FESST-cFEP, i.e. the cFEP of the ETN obtained from the application of FESST to the time series of the single distance C_β Gln_4−C_β Thr_{16} in Beta3s. The solid horizontal line indicates the barrier height of the 645-coords cFEP [9]. The dashed horizontal line displays the maximal barrier height found, which results when the heaviest 107 native nodes in the 645-coords cFEP are merged. The inset shows the FESST cFEPs with M= 3000 native nodes merged (black) and without merging (red).

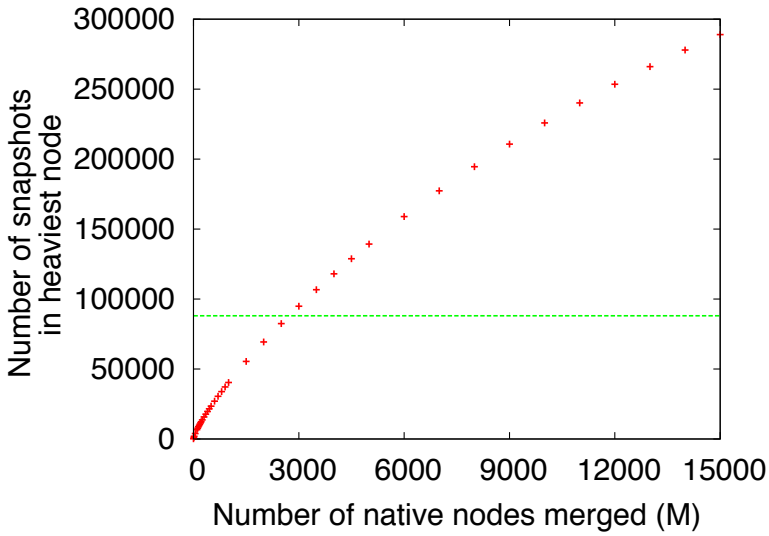

Figure S6: Number of time series bins (which corresponds to snapshots of the MD simulation) in the heaviest node of the FESST-ETN as a function of the number of merged nodes in the native basin. The window size is N = 100 and the acceptance cutoff is $\zeta = 0.3$. The horizontal line indicates the number of snapshots in the heaviest node of the 645-coordinates ETN.

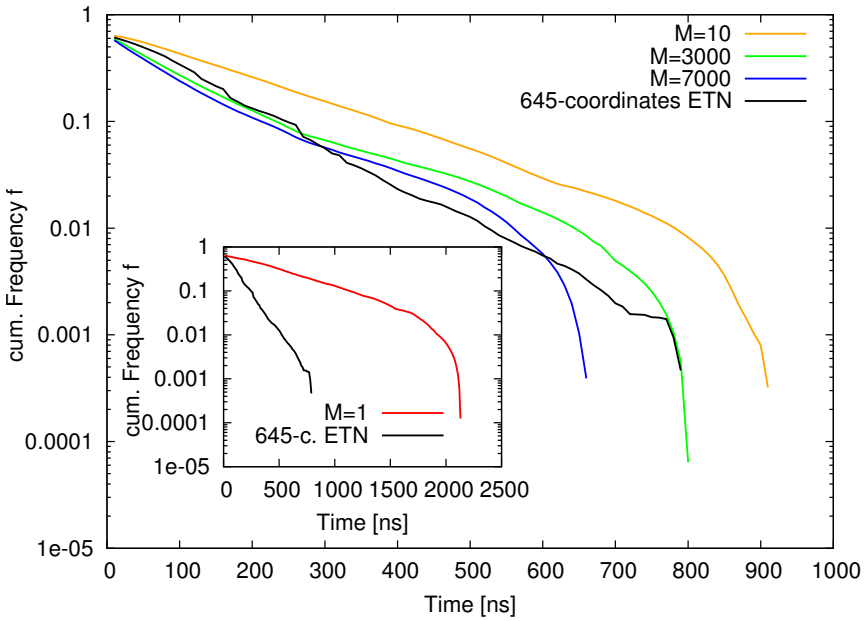

Figure S7: Comparison of folding kinetics for different representatives of the native basin. The plot shows the cumulative distribution of first passage times to the native node, $\mathrm{f}(t) = \int_t^\infty p(\tau)\mathrm{d}\tau$, where p is the probability distribution of the first passage time. All snapshots were used to calculate $\mathrm{f}(t)$. The plot for the 645-coordinates ETN (black lines) is shown as a reference [9]. The inset shows the much slower folding of the FESST-ETN without native node merging (M=1), which is a consequence of the small weight of the most populated node (Fig. S6).

1.4 Comparison of FESST performance for suboptimal intramolecular distances monitored

It is useful to evaluate the performance of FESST for a mediocre and bad separation of the native state peak from the rest of the basins. For this purpose the C_β-C_β distance between residues 1 and 13 (Fig. S8, left) and residues 8 and 18 (Fig. S8, right) are examined.

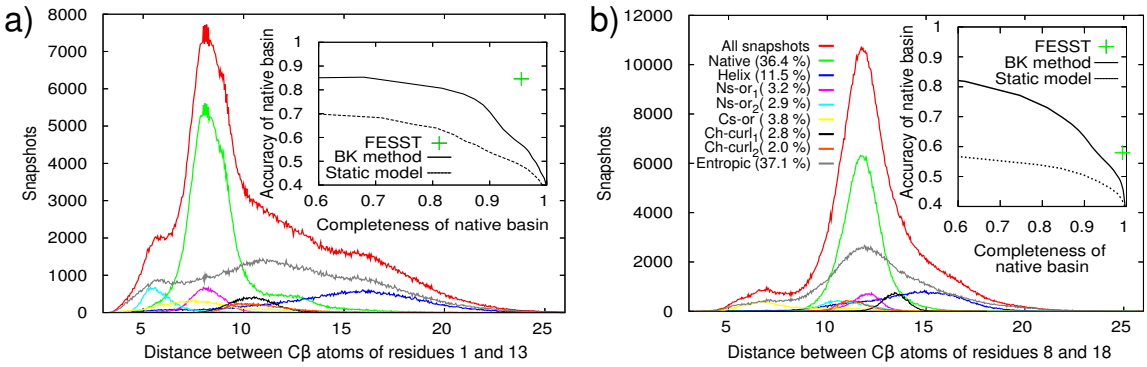

Figure S8: Distribution of inter-residue distances in different free-energy basins as identified using all 645 coordinates of Beta3s. The insets show the accuracy and completeness of the native state detection of FESST (cf. Fig. S3) compared to the BK procedure [10] and static model (described in the main text).

1.5 Choice of the window size for removal of basin overlap

Multiple free-energy basins may overlap on the cFEP because the short-time distribution of the distance is degenerate. To split the basin $\tilde{B}_2$ determined by FESST-cFEP as the first non-native basin (Fig. 2d), the long-time distribution of the distance for each time window $[t_2, t_2 + T]$ (the time bin t_2 belongs to $\tilde{B}_2$) is compared with the distribution of the distance in the entire native basin. The individual sub-basins correspond to individual peaks of the histogram of the comparison metric for a large range of window sizes T (Fig. S9). A value of T=40 ns is used in the main text.

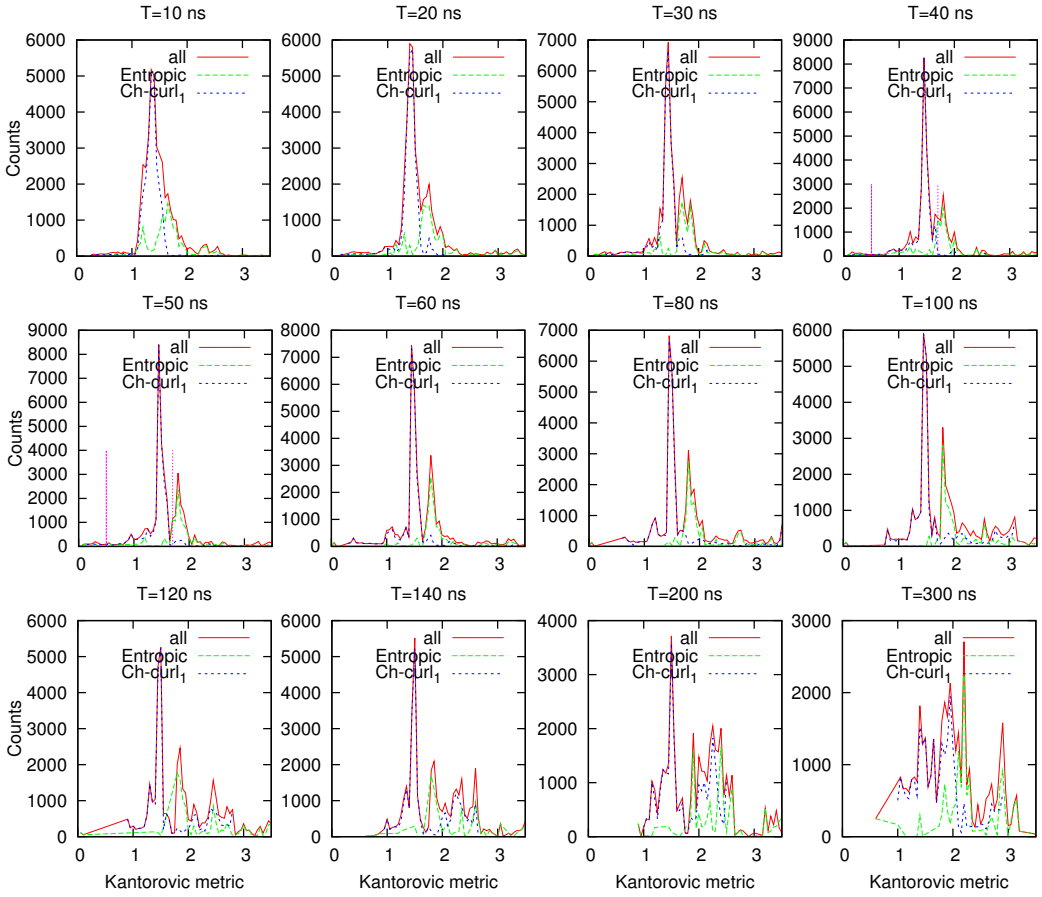

Figure S9: Effect of different window lengths in basin overlap removal. Compared is the Kantorovich metric distribution (area between cumulative histograms) between the long-time distance distributions in windows of varying length T around MD snapshots in $\tilde{B}_2$ and the native distance distribution, i.e., all MD snapshots in B_1 (see Fig. 2 for definition of $B_1, \tilde{B}_2$). This plot illustrates that the ranges of Kantorovich metric are different for the two different subbasins in $\tilde{B}_2$, i.e., Ch-curl$_1$ and Entropic.

1.6 Computational costs

Coarse-graining is the computational bottleneck, and the time it requires depends on the parameters used. Using a variant of the leader algorithm that preserves local kinetics (see main text) and the Kolmogorov-Smirnov test, coarse-graining of 10^6 time windows (of the distance between C_β-atoms of residues 4 and 16 with a window size of N = 100 and a cutoff parameter $\zeta = 0.3$) takes 6 hours on a recent XEON CPU with 2.33 GHz clock frequency. The CPU time depends on the acceptance cutoff. A cutoff of $\zeta = 0.38$ reduces the running time to 4.5 hours. The cFEP calculation takes only five to ten minutes. Very small memory requirements are needed for both procedures. Note that the determination of multiple free energy basins requires only one coarse-graining, but multiple cFEP calculations.

2 FESST application to an emulated FRET signal

To emulate a FRET experiment, the MD-generated time series of the distance between residues 4 and 16 in Beta3s is used together with a Markov state model to generate the photon time series (Fig. 5). The photophysical states considered are DA (both donor and acceptor in ground state), D^*A (donor in excited state, acceptor in ground state), DA^* and D^*A^*. The transition probabilities are approximated by the product of the transition rate and the time step (chosen to be $dt = 0.2$ ps, i.e., 100 observations along the MD saving interval of 20 ps). Finer time steps changed the photon counting results only marginally. For the intrinsic relaxation rates of donor and acceptor, $k_A = k_D = 2500\frac{1}{2\text{ ns}}$ is used. Direct excitation of the acceptor is set to 5% of the donor excitation rate. The Förster rate $k_F(r) = k_D \left(\frac{R_0}{r}\right)^6$ is calculated from the instantaneous distance r between the C_β-atoms of residues 4 and 16. The Förster radius is $R_0 = 12$ Å, which is the smallest radius that separates the distributions of FRET efficiencies of native and non-native conformations best. This separation is important, because there is an anticorrelation of the score of the native basin and the overlap of the distributions

in native and non-native state (data not shown).

The photon time series from the emulated FRET experiment is divided into bins of size S and the FRET efficiency $E_{\mathrm{FRET}} = \frac{n_{\mathrm{A}}}{n_{\mathrm{A}}+n_{\mathrm{D}}}$ is calculated for each bin (Fig. 5.C), where n_{A} and n_{D} are the number of acceptor and donor photons in the considered bin, respectively. The effect of the number of photons per bin is studied by the variation of the excitation rate (Fig. 6). For comparability with experiments, we report the number of photons emitted during the folding time. To improve the statistics at low emission rates, the photon time series is split into bins of S=0.4 ns, i.e., 20 MD snapshots. To improve sampling, detailed balance is imposed on the equilibrium transition network (ETN), i.e. the weight of each link is set to the average number of transitions in either direction. The score of the native state detection increases with the average emission rate, and reaches a plateau of 85% accuracy and 88% completeness (cf. Fig. S3 for definition) at an emission rate of about 5000 photons per folding time (Fig. 6), compared to 96% accuracy and 95% completeness obtained by applying FESST to the distance time series. The binning time S has little influence on the FESST performance (Fig. S10), whereas the length of the window T (parameter in FESST coarse-graining) shows a significant influence at low excitation rates (Fig. S11).

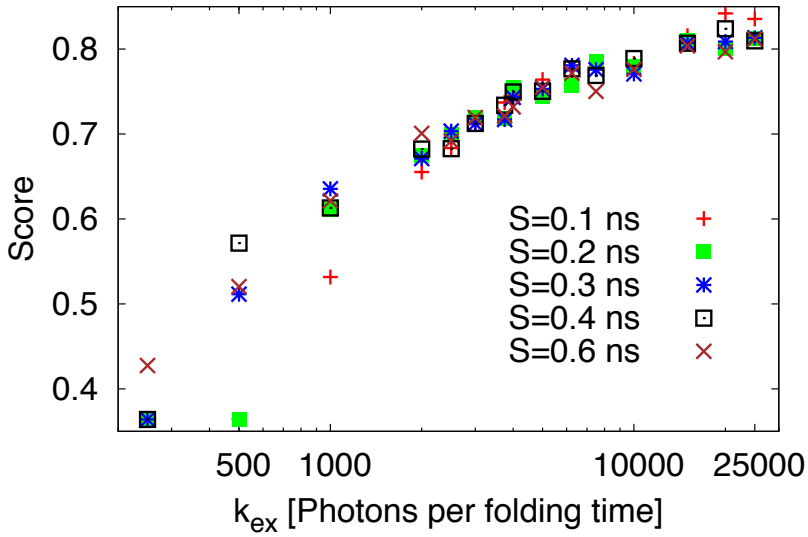

Figure S10: Robustness of the native basin detection in emulated FRET experiments upon change of binning time S. Note that the scores are calculated on a snapshot-wise basin assignment as for Fig. 6. For each excitation rate k_{ex}, only the highest score (tested are window sizes of $2, 4, 6, 8, 10, 20$, and 40 ns) is shown. The excitation rate k_{ex} is expressed as the number of photons emitted per folding time, which is 100 ns for Beta3s at 330 K (Fig. 6).

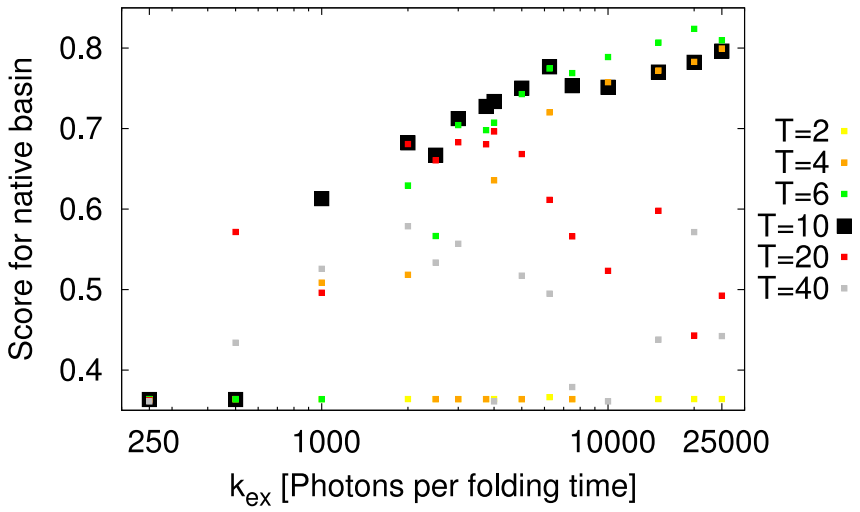

Figure S11: Effect of different window sizes T [ns] on FESST performance in emulated FRET experiments with 0.4 ns bins. The identification of the native basin is robust with respect to the choice of the window size in the range 6 ns $\leq$ T $\leq$ 10 ns. The same setup as for the data shown in Fig. 6 and described in the SI is used. As in Fig. 6, the excitation rate k_{ex} is expressed as the number of photons emitted during the folding time, which is 100 ns for Beta3s at 330 K [11].

3 One-dimensional two-state system: Resolution limit of FESST

It is useful to investigate the resolution limit of FESST using a simple model (Fig. S12). The time evolution of the monitored signal is given by Langevin dynamics of a particle in a one-dimensional potential. To model a two-state system, the potential is switched with a constant rate between two harmonic wells (Fig. S12 a). The sequence of emitted photons is determined by a Gillespie-type simulation [12]. The time series of FRET efficiencies is derived from the binned photon sequence and analyzed by FESST as described for Beta3s with detailed balance imposed. Remarkably, the accuracy of FESST is always higher than the static model even for very small separations of the minima (Fig. S12 b,c,d). FESST can discriminate the minima based on the curvature alone, albeit with a relatively large number of detected photons required (cf. Fig. S12 and Fig. S13). Although the simple model does not reflect the complexity of a multidimensional system, the present

results indicate that the reliable operation of FESST requires the detection of 100 to 1000 photons during the residence time in one free-energy state. In real single-molecule FRET experiments, 100 to 1000 photons can be detected in about one to ten milliseconds. Accordingly, we expect FESST to be a suitable approach for determining the properties of free energy surfaces of molecules that exhibit dynamics in this range or slower, thus covering a large part of the biologically important time scales [13, 14, 15, 16].

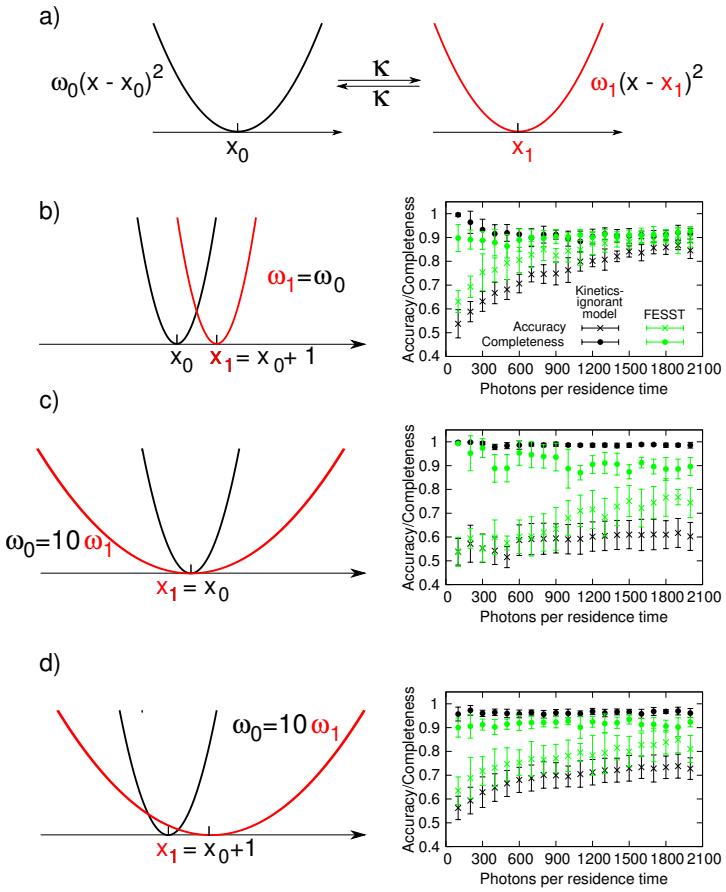

Figure S12: Resolution limits of FESST examined with a one-dimensional two-state model. (a) Schematic illustration of the model: Dynamics of a massive particle in a harmonic potential that switches from one shape and position to the other with rate κ. The position time series is then transformed to a time series of FRET efficiencies as for Beta3s (the equilibrium position is defined as $x_0 = 10$ A.U. (arbitrary units), the Förster radius as $R_0 = 12$ A.U., the curvature of potential 0 as $\omega_0 = 100$ A.U., and the interchange rate is assumed to be $\kappa = 0.005$ A.U.). (b-d, left) Illustration of the potential's shape and position. (b-d, right) Accuracy and completeness of the basin with index 0 as detected by FESST and the best static model in ten independent simulations for the potential's setup on the left. The parameters for FESST are optimized using the height of the unfolding barrier determined self-consistently (cf. Fig. 2c). For the static model, the basin is formed by all bins with a FRET efficiency lower than a given cutoff. To make the most stringent comparison, the detection quality of the static model is maximized by finetuning both the length of the binning interval and the cutoff value based on the knowledge of the solution. The shapes and positions of the potentials used to investigate the different scenaria are given in the individual panels: (b) examines the effect of a shift, (c) the effect of a broadened potential with identical equilibrium position, and for (d) both curvatu./Paper/leader-single-molecule/vorschlag-FRET/re and equilibrium position are changed. For a minor shift $x_1 - x_2 = 1$ A.U. of the two minima of the potential (panels (c) and (d)), FESST yields significantly more accurate solutions with only slightly lower completeness than the static model for as few as 200 photons per residence time. An increased amount of about 1000 photons per residence time is required if the equilibrium positions match and the curvatures differ by a factor of ten (panel c).

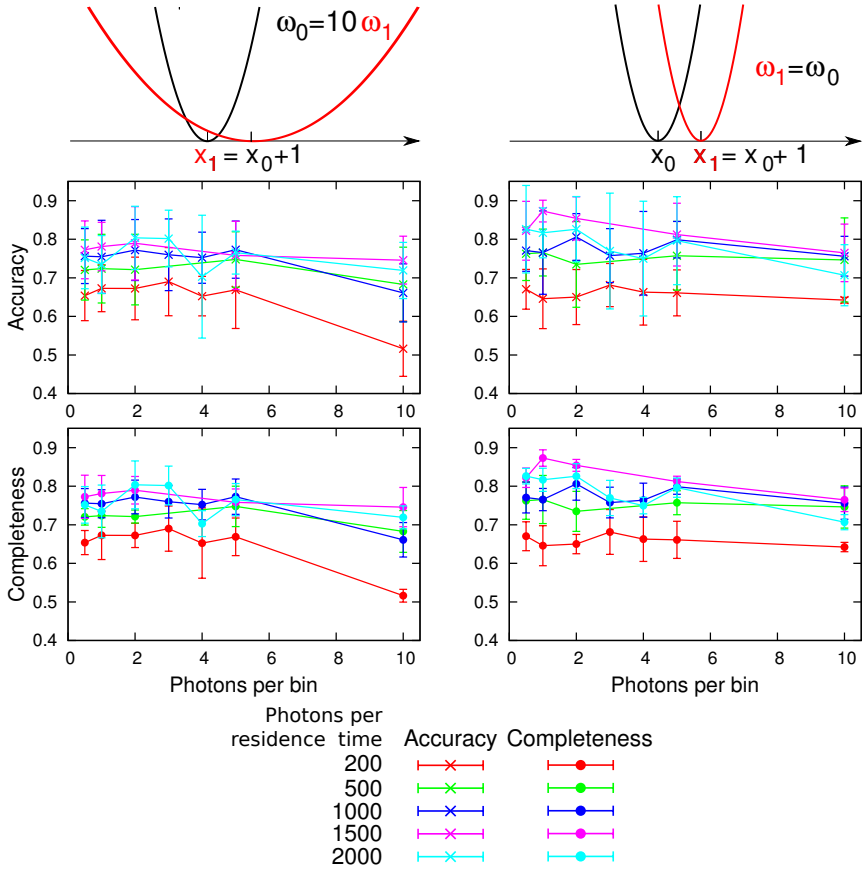

Figure S13: Dependence of FESST performance in the one-dimensional two-state model on the number of photons per FRET bin. The accuracy and completeness of the basin with index 0 (cf. Fig. S12) as detected by FESST depend mainly on the number of photons emitted during the residence time, and only weakly on the explicit number of photons per bin.

4 FESST application to real experimental data (single-molecule FRET on λ-repressor)

4.1 Expression, purification, and labeling of λ-repressor

A plasmid (pET-47b(+)-vector, Novagen) containing the custom-synthesized and codon-optimized sequence coding for the monomeric N-terminal fragment of λ-repressor including an amino terminal hexahistinie tag was purchased from Celtek Genes (Nashville, USA). Threonine 8 and lysine 70 were replaced by cysteine residues for fluorophore labeling. The final amino acid sequence was GPSLCQEQLEDARRLKAIYEKKKNELGLSQESVADKMGMGQSGVGALFNGINALNAYNAALLAKILCVSVEEFSPSIAREIR. The protein was expressed in E.coli BL21 cells at 37°C in LB medium containing kanamycin and 1mM IPTG for induction. The resulting inclusion bodies were harvested by centrifugation after cell lysis with a French pressure cell. Resolubilized protein was subjected to immobilized metal ion affinity chromatography (IMAC; HisTrap H, GE Healthcare) at pH 8. The single peak eluting in the imidazole gradient was collected. The N-terminal His-Tag was cleaved with HRV 3C protease. Uncleaved λ-repressor and protease were separated using IMAC and gel filtration (Superdex 75, GE Healthcare). Labeling was performed essentially as described previously [17]. Purified protein was reacted first with Alexa Fluor 488 maleimide (Invitrogen) at substochiometric concentrations according to the supplier's recommendations. The resulting products were separated using anion exchange chromatography (MonoQ 5/50 GL, GE Healthcare). Singly labeled protein was then concentrated by ultrafiltration (Centricon, Millipore) and reacted with an excess of Alexa Fluor 594 maleimide. Monomeric λ-repressor with one donor and one acceptor dye was purified by anion exchange and size exclusion chromatography. All steps were carried out at high concentrations of denaturant. Correct labeling was confirmed by electrospray ionization mass spectroscopy.

4.2 Single molecule spectroscopy

Measurements were carried out in a MicroTime200 (PicoQuant, Berlin, Germany) with continuous wave laser excitation at 488 nm essentially as described previously [17, 18]. In measurements for the FESST analysis, protein was refolded and diluted from a stock in 8 M guanidinium chloride into a buffer containing 50 mM sodium phosphate buffer pH 7, 0.01% Tween, 150 mM beta-mercaptoethanol, 10 mM cysteamine to a final concentration of guanidinium chloride of 0.68 M and 16 pM of protein. The sample was measured in a temperature-controlled cell [18] with a temperature of 12°C at the position of the confocal volume. The laser power was adjusted to 600 μW. Nanosecond correlation measurements were performed at a protein concentration of 1 nM with a laser power of 30 μW and analyzed as described previously [19].

4.3 Treatment of photon time series with bursts

As freely diffusing molecules are observed in our FRET-experiments, the photon time series is structured in bursts. This signal has to be converted into a time series of nodes by coarse-graining with FESST (Fig. 5.C, S14).

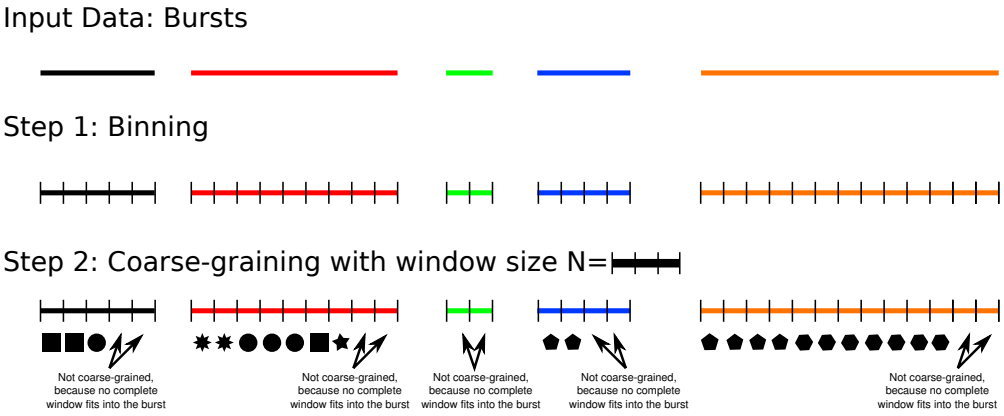

Figure S14: FESST coarse-graining for photon time series from individual bursts. As for a continuous photon time series, the data are first binned (Step 1). In the second step, bins are assigned to a node if their short-time window fits completely in the burst.

4.4 Robustness of FESST upon variation of coarse-graining parameters

This analysis shows that the determination of the folded and unfolded populations/basins is robust for $25 \leq N \leq 30$ (Fig. S15, left) and $0.3 \leq \zeta \leq 0.4$ (Fig. S15, right).

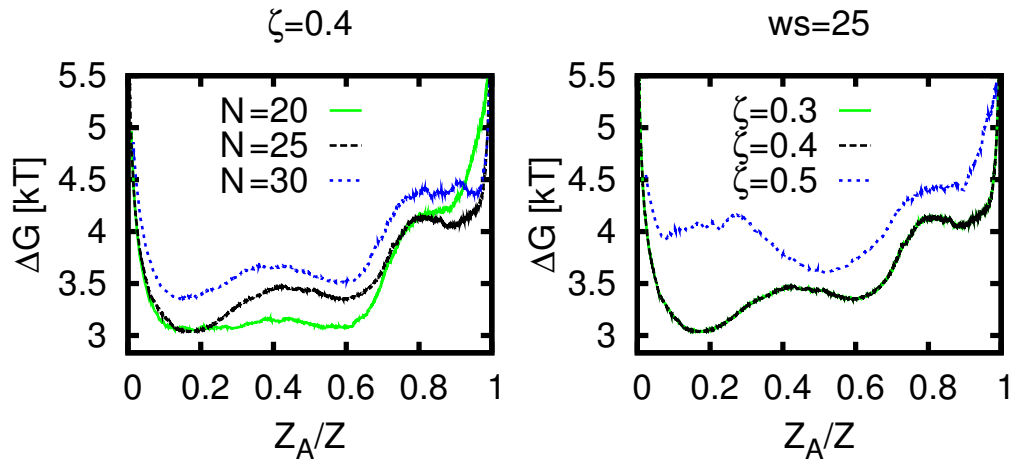

Figure S15: Differences in cut-based free-energy profiles (cFEP) upon changes of the FESST coarse-graining parameters. The input signal consists of a photon time series measured in a FRET experiment of freely diffusing λ-repressor fragments. Bursts are characterised by at most $\delta t = 70\mu$s between two successive photons. FESST is applied to the time series of FRET efficiencies calculated for 0.1 ms bins. The highest barrier resulted for a window size N = 25 bins (which corresponds to a time $\tau = 2.5$ ms) and an acceptance cutoff $\zeta = 0.4$. Note that the $\zeta = 0.3$ (green) and $\zeta = 0.4$ (black) curves overlap fully (right panel), and thus the former is not visible.

4.5 cFEP with the folded state as a reference

It is interesting to compare the cFEP obtained using as a reference the most populated node, which is a representative of the unfolded state (Fig. 7b), with the cFEP from the most populated node in the folded state (Fig. S16). The two cFEPs are consistent. In particular, the (un)folding barrier height and the two main basins are similar.

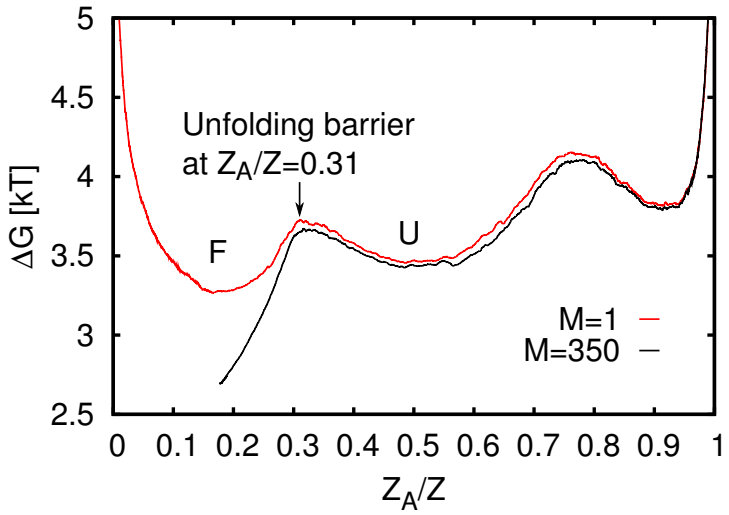

Figure S16: Cut-based free-energy profile (cFEP) from the folded state. The location of the barrier for unfolding is indicated (black arrow). A significantly higher unfolding barrier (black curve) is obtained by merging the M=350 most populated nodes in the folded basin as determined with M=1 (red curve).

4.6 Imposing detailed balance on the ETN of lambda-repressor

For the lambda-repressor, detailed balance (DB) is not imposed. The cFEPs for the ETNs with or without DB differ as shown in S17. The cFEP for the ETN

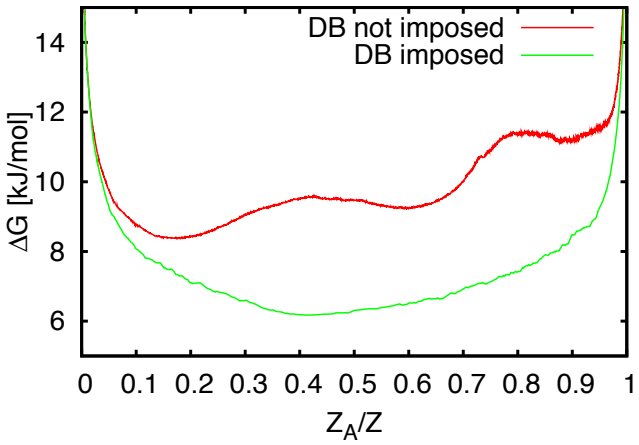

Figure S17: Comparison of the cut-based free-energy profile (cFEP) from the most populated state with and without detailed balance (DB) imposed. The cFEP from the ETN with DB imposed has no barrier in contrast to the cFEP from the unmodified ETN.

with detailed balance imposed shows no barriers. The reason for this observation are spurious transitions observed only in individual bursts. In the ETN, these transitions show up as chains visited only once and only in one direction (a chain is a sequence of nodes $A_1, \ldots, A_n$, where A_i is linked exclusively to A_{i+1} for $i = 1, \ldots, n-1$). This hypothesis is confirmed by comparing the cFEP of the ETN without DB to the cFEP of ETN (cf. Fig. S18) in which on all links but those belonging to a chain visited just once are symmetrised (the weight of the link is set to the average number of transitions in either direction).

As a second test, all links in a unidirectional chain are removed (blue cFEP in S18). The position of the first barrier is equal for all three profiles. The barrier for the cFEP from the ETN with links removed (blue curve in S18) is higher, because spurious transitions are removed. These unidirectional chains are present, because actual states of the system are split into multiple nodes due to shot noise.

As illustrated above, imposing detailed balance on a system such as the lambda-

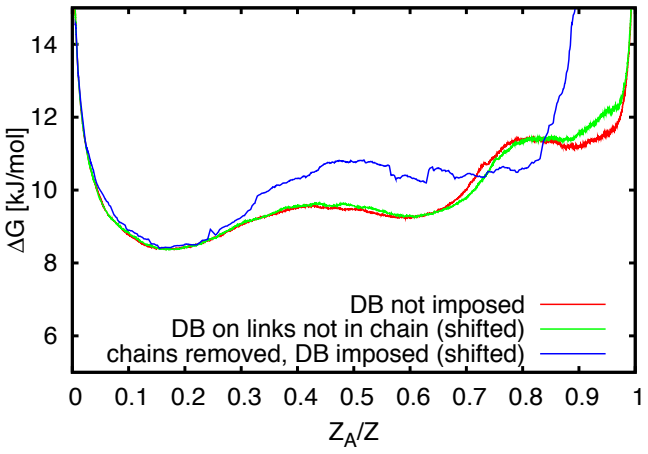

Figure S18: Assessment of the effect of unidirectional chains on the cut-based free-energy profile (cFEP) from the most populated state when detailed balance (DB) is imposed. All links apart from those in a unidirectional chain can be symmetrised (i.e. the weight of the link is set to the average number of transitions in either direction) without any change on the cFEP. Removing the chains from the cFEP leaves the position of the barrier invariant. The height of the barrier increases, because spurious transitions are removed. Two of the three cFEPs were shifted along the y-axis to bring the bottom of the first basin on the left to the same reference value.

repressor leads to a different ETN. A random walker on a unidirectional chain walks from start to end with shortest number of steps. If a chain is symmetrised, the random walker diffuses for artificially long times along the symmetrised chain and might even return to its beginning, which is a prediction in total contradiction to the experimental results. However, if detailed balance is not imposed, the values for the flux between the large nodes (number of transitions between them) are correct.

References

[1] De Alba E, Santoro J, Rico M, Jiménez MA (1999) De novo design of a monomeric three-stranded antiparallel β-sheet. *Protein Science* 8:854–865.

[2] Ferrara P, Caflisch A (2000) Folding simulations of a three-stranded antiparallel β-sheet peptide. *Proc. Natl. Acad. Sci. USA*. 97:10780–10785.

[3] Brooks BR, et al. (1983) CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* 4:187–217.

[4] Brooks BR, et al. (2009) CHARMM: The biomolecular simulation program. *J. Comput. Chem.* 30:1545–1614.

[5] Ferrara P, Apostolakis J, Caflisch A (2002) Evaluation of a fast implicit solvent model for molecular dynamics simulations. *Proteins: Structure, Function, and Bioinformatics* 46: 24–33.

[6] Neria E, Fischer S, Karplus M (1996) Simulation of activation free energies in molecular systems. *J. Chem. Phys.* 105:1902–1921.

[7] Cavalli A, Ferrara P, Caflisch A (2002) Weak temperature dependence of the free energy surface and folding pathways of structured peptides. *Proteins: Structure, Function, and Bioinformatics* 47: 305–314.

[8] Krivov SV, Karplus M (2006) One-dimensional free-energy profiles of complex systems: Progress variables that preserve the barriers. *J. Phys. Chem. B* 110:12689–12698.

[9] Krivov SV, Muff S, Caflisch A, Karplus M (2008) One-Dimensional Barrier Preserving Free-Energy Projections of a beta-sheet Miniprotein: New Insights into the Folding Process. *J. Phys. Chem. B* 112:8701–8714.

[10] Baba A, Komatsuzaki T (2007) Construction of effective free energy landscape from single-molecule time series. *Proc. Natl. Acad. Sci. USA*. 104:19297–19302.

[11] Muff S, Caflisch A (2008) Kinetic analysis of molecular dynamics simulationsa reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a β-sheet miniprotein. *Proteins: Structure, Function, and Bioinformatics* 70: 1185–1195.

[12] Gopich IV, Szabo A (2009) Decoding the Pattern of Photon Colors in Single-Molecule FRET. *J. Phys. Chem. B* 113:10965–10973.

[13] Joo C, et al. (2006) Real-Time Observation of RecA Filament Dynamics with Single Monomer Resolution. *Cell* 126:515–527.

[14] Joo C, Balci H, Ishitsuka Y, Buranachai C, Ha T (2008) Advances in Single-Molecule Fluorescence Methods for Molecular Biology. *Annual Review of Biochemistry* 77:51–76.

[15] Borgia A, Williams PM, Clarke J (2008) Single-Molecule Studies of Protein Folding. *Annual Review of Biochemistry* 77:101–125.

[16] Chung HS, Louis JM, Eaton WA (2009) Experimental determination of upper bound for transition path times in protein folding from single-molecule photon-by-photon trajectories. *Proc. Natl. Acad. Sci. USA.* 106:11837–11844.

[17] Hoffmann A, Kane A, Nettels D, Hertzog DE, Baumgärtel P, Lengefeld J, Reichardt G, Horsley DA, Seckler R, Bakajin O, Schuler B (2007) Mapping protein collapse with single-molecule fluorescence and kinetic synchrotron radiation circular dichroism spectroscopy. *Proc. Natl. Acad. Sci. USA.* 104:105–110.

[18] Nettels D, Müller-Späth S, Küster F, Hofmann H, Haenni D, Rüegger S, Reymond L, Hoffmann A, Kubelka J, Heinz B, Gast K, Best RB, Schuler B (2009) Single-molecule spectroscopy of the temperature-induced collapse of unfolded proteins. *Proc. Natl. Acad. Sci. USA.* 106:20740-20745.

[19] Nettels D, Gopich IV, Hoffmann A, Schuler B (2007) Ultrafast dynamics of protein collapse from single-molecule photon statistics. *Proc. Natl. Acad. Sci. USA.* 104:2655–2660.

Concluding Remarks

In this work, primarily on λ repressor, we made steps toward a key goal in the field of protein biophysics: the "molecular movie", a series of high resolution temporal and spatial data describing protein structure and dynamics in all states over a wide range of timescales.

So far, a detailed molecular movie has only become fully available within the framework of molecular dynamics simulations (Pande *et al.*, 2003; Rao and Caflisch, 2004; Bowman *et al.*, 2011; Lindorff-Larsen *et al.*, 2011). The direct observation of protein folding transition states and and the full description of the folding reaction landscape has so far only been possible *in silico*. This has made the study of fast processes of ultra-fast protein folding and dynamics in the unfolded state especially compelling, since by using data obtained experimentally one will eventually be able to calibrate and cross-validate simulations. Several methods, such as time-resolved infra-red spectroscopy (Ihalainen *et al.*, 2007), picosecond Laue crystallography (Schotte *et al.*, 2003), laser temperature-jump relaxation(Yang and Gruebele, 2003) or dynamic NMR (Arora *et al.*, 2004), can access interesting timescales relevant to fast protein folding. FRET-based methods can contribute high temporal resolution and the ability to easily employ single-molecule high-throughput setups. Maybe most importantly single-molecule FRET holds the promise to be able to explore the whole free energy landscape of the folding reaction and possibly reveal the whole distribution of reaction pathways.

The study of protein folding and the properties of the unfolded state are connected by the fact that he unfolded state is the educt of the folding reaction. Its dynamics and dimensions will dictate the kinetics of the folding, as they will limit the time it takes for two distant parts of a chain to form a long range interaction. Recently Soranno *et al.* (Soranno *et al.*, 2012) could quantify the internal friction in unfolded/intrinsically disordered proteins, finding stronger internal friction in more collapsed chains. It would be interesting how this will change at higher temperatures. The exact molecular origin of this internal friction is not entirely clear, but might involve the solvation of the unfolded state. Possibly, internal friction can be used as a reporter on the amount of solvent in the unfolded state. To elucidate if clustering of hydrophobic residues is the reason for the unusual behavior of λ repressor, I would propose to investigate polypeptides of a repeating residue pattern. By enriching and varying hydrophobic residues it might be possible to recreate this behavior in a more controlled environment, sim-

ilar to the experiments by Walker *et al.* (Li and Walker, 2011). For example, series of experiments on repeats of alanine-alanine-leucine vs. alanine-leucine vs alanine-leucine-leucine etc. could report on the minimum size of clusters needed to affect solvation behavior and the minimum distance between hydrophobic residues to form such clusters (Murnen *et al.*, 2012). Ultimately one needs to investigate the water molecules. Since this is very hard in a single molecule context, molecular dynamic simulations are maybe a viable road to go. The water network around hydrophobic residues already has been investigated and found to have some temperate dependent roll-over behavior (Oleinikova *et al.*, 2010).

Internal friction also is interesting in the context of fast protein folding. It has been found that the degree of internal friction in the transition state, in special cases, can be higher that previously thought (Wensley *et al.*, 2010). This should have a profound effect on the barrier transition path times. A choice of model protein which exhibits a high degree of internal friction but still has a high folding rate, such as the spectrin domains, might be a good candidate to resolve barrier transitions and get one step closer to a ”molecular movie” of the folding process.

Bibliography

Arora P.; Oas T.G.; Myers J.K. Fast and faster: a designed variant of the B-domain of protein A folds in 3 microsec. *Protein science : a publication of the Protein Society*, **13**(4):847–853 (2004).

Bowman G.R.; Voelz V.A.; Pande V.S. Atomistic folding simulations of the five-helix bundle protein λ(6–85). *Journal of the American Chemical Society*, **133**(4):664–667 (2011).

Ihalainen J.A.; Bredenbeck J.; Pfister R.; Helbing J.; Chi L.; van Stokkum I.H.M.; Woolley G.A.; Hamm P. Folding and unfolding of a photoswitchable peptide from picoseconds to microseconds. *Proceedings of the National Academy of Sciences of the United States of America*, **104**(13):5383–5388 (2007).

Li I.T.S.; Walker G.C. Signature of hydrophobic hydration in a single polymer. *Proceedings of the National Academy of Sciences of the United States of America* (2011).

Lindorff-Larsen K.; Piana S.; Dror R.O.; Shaw D.E. How fast-folding proteins fold. *Science*, **334**(6055):517–520 (2011).

Murnen H.K.; Khokhlov A.R.; Khalatur P.G.; Segalman R.A.; Zuckermann R.N. Impact of Hydrophobic Sequence Patterning on the Coil-to-Globule Transition of Protein-like Polymers. *Macromolecules*, **45**(12):5229–5236 (2012).

Oleinikova A.; Brovchenko I.; Singh G. The temperature dependence of the heat capacity of hydration water near biosurfaces from molecular simulations. *Epl*, **90**(3):36001 (2010).

Pande V.S.; Baker I.; Chapman J.; Elmer S.P.; Khaliq S.; Larson S.M.; Rhee Y.M.; Shirts M.R.; Snow C.D.; Sorin E.J.; Zagrovic B. Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing. *Biopolymers*, **68**(1):91–109 (2003).

Rao F.; Caflisch A. The protein folding network. *Journal of Molecular Biology*, **342**(1):299–306 (2004).

Schotte F.; Lim M.; Jackson T.A.; Smirnov A.V.; Soman J.; Olson J.S.; Phillips G.N.; Wulff M.; Anfinrud P.A. Watching a protein as it functions with 150-ps time-resolved x-ray crystallography. *Science*, **300**(5627):1944–1947 (2003).

Soranno A.; Buchli B.; Nettels D.; Cheng R.R.; Müller-Späth S.; Pfeil S.H.; Hoffmann A.; Lipman E.A.; Makarov D.E.; Schuler B. Quantifying internal friction in unfolded and intrinsically disordered proteins with single-molecule spectroscopy. *Proceedings of the National Academy of Sciences of the United States of America* (2012).

Wensley B.G.; Batey S.; Bone F.A.C.; Chan Z.M.; Tumelty N.R.; Steward A.; Kwa L.G.; Borgia A.; Clarke J. Experimental evidence for a frustrated energy landscape in a three-helix-bundle protein family. *Nature*, **463**(7281):685–688 (2010).

Yang W.Y.; Gruebele M. Folding at the speed limit. *Nature*, **423**(6936):193–197 (2003).

Lebenslauf

René Wuttke

Persönliche Daten

18.06/1979	geboren in Potsdam, Deutschland

Ausbildung

08/92–06/99	Abitur am Hermann-von-Helmholtz-Gymnasium mit Erreichen der allgemeinen Hochschulreife
10/00–08/05	Studium der Biochemie an der Universität Potsdam mit Erlangen des Diploms im Fach Biochemie
05/04–01/05	Diplomarbeit an der Carnegie Institution of Washington, Stanford, USA zum Thema "Biochemical and genetic analysis of mechanisms involved in cell wall formation in *Arabidopsis*" unter Betreuung von Prof. Dr. L. Willmitzer und Prof. Dr. C. Somerville
seit 09/05	Promotion als Doktorand am Biochemischen Institut der Universität Zürich unter der Leitung von Prof. Dr. Ben Schuler

Acknowledgements

Several people contributed to this work and supported me throughout the years.

I like to thank Prof. Ben Schuler, for giving me the opportunity to work on this project, introducing me to exciting field of single molecule protein folding, proving excellent research facilities and his constant support.

I like to thank Prof. Raimund Dutzler and Prof. Peter Hamm for being on my PhD committee.

I like to thank Dr. Frank Hillger and Dr. Hagen Hofmann for numerous discussion and mentorship, on techniques and theories new to me.

I like to thank Dr. Phillip Schütz for giving me the opportunity to contribute to the work on the FESST method and Bengt Wunderlich for collaborative efforts the fast microfluidic mixer.

I like to thank all past and current members of the Schuler group - Armin, Sonja, Daniel, Frank & Frank, Dominik, Luc, Bengt, Hagen, Philipp, Andrea & Andrea, Ruth, Jennifer, Stephan, Franziska, Alessandro, Madeleine - not only for their help and scientific input, but also for the great working atmosphere I could enjoy over the years.

I cordially thank my family and friends for their constant support and interest. Especially I like to thank my parents, Margrit and Werner, for giving me the possibility to study, for their encouragement, and their ongoing support.